Maria da Conceição Nunes Elói Veiga de Almeida

Licenciada em Gestão

A Aplicação da Teoria de Valores Extremos ao Tráfego da Ponte 25 de Abril

Dissertação para obtenção do Grau de Mestre em

Matemática e Aplicações ramo Matemática Financeira

Orientador: Frederico Almeida Gião Gonçalves Caeiro,
Professor Auxiliar, Universidade Nova de Lisboa

Júri

Presidente: Professor Doutor Filipe José Gonçalves Pereira Marques
Arguente: Professora Doutora Dora Susana Raposo Prata Gomes
Vogal: Professor Doutor Frederico Almeida Gião Gonçalves Caeiro

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Março, 2019

A Aplicação da Teoria de Valores Extremos ao Tráfego da Ponte 25 de Abril

Este documento foi gerado utilizando o processador (pdf)LaTeX, com base no template "novathesis" [1] desenvolvido no Dep. Informática da FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

À minha família

Agradecimentos

Em primeiro lugar, quero agradecer ao Professor Doutor Frederico Caeiro que me orientou ao longo deste processo! Pela sua disponibilidade constante, pelos e-mails sempre respondidos, pelo esclarecimento de dúvidas por mais "ilógico" que fosse, pelo facto de sempre me dar tanta liberdade para explorar uma área do seu interesse ao meu gosto, pela sua visão de conjunto tão precisa e prática, pelos conselhos e comentários tão úteis e pela proposta deste estudo que acabou por me entusiasmar tanto.

Quero agradecer à Faculdade de Ciências e Tecnologia pelos recursos disponibilizados para a elaboração desta dissertação, por todas as ferramentas e conhecimentos necessários para a elaboração do meu trabalho.

Também quero agradecer muito em particular ao Exmo. Sr. Dr. Rui Cesar Ilha Luso Soares do Instituto da Mobilidade e dos Transportes, I.P.! Pela incansável disponibilidade! Quando tantas portas me fecharam e me abriu uma tão grande janela! Pelo incontável número de emails que respondeu, pela chamada que atendeu, pelo esclarecimento de dúvidas de qualquer tipo, pelos dados variados que foram pedidos, e não só! É difícil expressar em palavras o enorme agradecimento que lhe devo! Muito Obrigada!

Depois tenho que agradecer imenso a toda a minha família! Principalmente, aos meus pais e irmãos! Pelos conselhos sábios do meu pai, pelo apoio incondicional da minha mãe, pela ajuda na clarificação de "foco" da Mariana, pelo tão necessário sentido de humor do Miguel e pela escuta tão atenta do João! Não deixando de agradecer de forma especial ao meu avô, à minha tia Coim e aos meus tios. Nunca me faltou o vosso apoio e palavras animadoras com as quais sempre se trabalhava com melhor e maior ânimo!

Um muito obrigada aos meus colegas de Mestrado! E a tantas pessoas amigas! Pela escuta atenta e interessada, conselhos tão bons, por todo o apoio e não só! Um obrigada especial à Ana pela disponibilidade de me "acolher" quando precisava de me focar mais nos estudos, por me tirar as dúvidas sobre o LaTeX e por todos os seus conselhos tão práticos e úteis. Um obrigada à Raquel por me ouvir tantas e tantas vezes, pelos conselhos da experiência "das teses", por sempre me dar "na cabeça" quando necessário e por me apoiar! Não cabem os nomes de todos mas não posso (nem quero) deixar de referir um obrigado especial à Teresa, à São e ao Salvador!

E a todos e a cada um: Muito Obrigada por tudo!

Resumo

A Teoria dos Valores Extremos permite o estudo dos acontecimentos extremos que são possivelmente desastrosos e de grande impacto para a sociedade. O comportamento dos Extremos pode ser modelado por uma das três distribuições – Gumbel, Fréchet e Weibull – se bem que estas distribuições podem ser representadas numa única expressão, a distribuição Generalizada de Valores Extremos (GEV).

Nesta dissertação, serão analisados os números de veículos que atravessam a Ponte 25 de Abril, nos dois sentidos, diariamente. Também serão efetuadas duas análises, consequentes destes dados, uma com base na sazonalidade e outra relativa ao valor das receitas das portagens cobradas na travessia desta Ponte.

Será utilizada uma abordagem paramétrica para a inferência estatística sobre acontecimentos raros. Para isso serão utilizados três modelos: o Modelo GEV (também conhecido como Modelo dos Máximos Anuais), o Modelo GEV Multivariado (ou Modelo Estatístico das r maiores observações) e o Modelo Generalizado de Pareto (GP ou Modelo dos excessos acima do limiar). Estes modelos são muito usados em diversas áreas.

Nesta tese é feita uma descrição do fluxo de tráfego na Ponte 25 de Abril e os Métodos dos Valores Extremos são utilizados para fazer uma previsão do comportamento desse mesmo tráfego. Serão estimados níveis de retorno, períodos de retorno e probabilidades de excedência. Será utilizado o Método da Máxima Verosimilhança para a estimação de parâmetros e o Método do perfil Log-Verosimilhança para a estimação de Intervalos de Confiança.

Palavras-chave: Teoria dos Valores Extremos, Modelo GEV, Modelo GEV Multivariado, Modelo Generalizado de Pareto (GP), Método da Máxima Verosimilhança, Tráfego Rodoviário.

Abstract

The Extreme Values Theory enables the study of extreme events that are possibly disastrous and of great impact for society. The behaviour of the Extremes can be modelled by using one of three distributions – Gumbel, Fréchet and Weibull – even though they can be represented in a single expression, the Generalized Extreme-Value distribution (GEV).

In this dissertation, the numbers of vehicles crossing daily and in both directions in the 25 de Abril Bridge will be analysed. Two analyzes will also be carried out, as a result of these data, one based on the verified seasonality and another in relation to the tolls and revenues collected in the crossing of this Bridge.

A parametric approach will be used for statistical inference about rare events. To achieve this three methods will be used: the GEV Model (also known as the Annual Maximum Model), the Multivariate GEV Model (or r Largest order statistic Model) and the Generalized Pareto Model (GP or Peak Over Threshold Model). These models are widely used in various areas.

In this thesis a description is made of the traffic flow in the 25 de Abril Bridge and the Methods of the Extreme Values are used to make a prediction of the behavior of this traffic. Return levels, return periods and probabilities of exceedance will be estimated. The Maximum Likelihood Method will be used for the estimation of parameters and so will the Profile Log-Likelihood Method when estimating Confidence Intervals.

Keywords: Extreme Values Theory, GEV Model, Multivariate GEV Model, Generalized Pareto Model, Maximum Likelihood Method, Road Traffic.

Índice

Lista de Figuras

Lista de Tabelas

Siglas

f.d.	Função de distribuição.
f.d.p.	Função de densidade de probabilidade.
GEV	Generalizada de Valores Extremos (em inglês, *Generalized Extreme Value*).
GP	Generalizada de Pareto.
GVRM	Gráfico ou Gráficos de Vida Residual Média.
i.i.d.	Independentes e identicamente distribuídas.
IC	Intervalo ou Intervalos de Confiança.
MV	Máxima Verosimilhança.
NR	Nível ou Níveis de Retorno.
TLC	Teorema Limite Central.
v.a.	variável aleatória.
v.a.'s	variáveis aleatórias.

CAPÍTULO 1

Introdução

A Teoria dos Valores Extremos proporciona técnicas de inferência estatística orientada para o estudo de comportamentos estocásticos extremais.

Esta teoria é frequentemente utilizada para obter distribuições de probabilidade do máximo ou mínimo de variáveis aleatórias (v.a.'s) independentes e identicamente distribuídas (i.i.d.), bem como para modelar a distribuição de excessos acima de um certo nível. Por exemplo, supondo que a sucessão de v.a.'s i.i.d. $X_1, X_2, \ldots$ representa o número de veículos que passam diariamente numa ponte, poder-se-á então estudar a distribuição da v.a.

$$M_n = \max\{X_1, \ldots, X_n\}$$

que representa o valor máximo diário de veículos durante um período de n observações. Se $F(x)$ representa a função distribuição (f.d.) de X_i então a função distribuição de M_n é

$$\begin{aligned}\Pr(M_n \leq x) = \Pr(X_1 \leq x \cap X_2 \leq x \cap \cdots \cap X_n \leq x) = \Pr(X_1 \leq x)\Pr(X_2 \leq x)\ldots\Pr(X_n \leq x) = \\ = F(x)F(x)\ldots F(x) = [F(x)]^n.\end{aligned}$$

Como geralmente $F(x)$ é desconhecida, para se obter a distribuição de M_n, recorre-se à teoria assintótica de valores extremos, que teve o seu início com os trabalhos de Fréchet, Fisher e Tippett (1928), Mises (1936) e o seu auge foi com o trabalho de Gnedenko (1943) que obteve as condições necessárias e suficientes que garantem a existência de um dos três tipos de distribuição limite para o máximo de v.a.'s i.i.d., nomeadamente a distribuição de Gumbel, Fréchet e Weibull.

Esta teoria é de suma importância para conhecer o comportamento de valores excessivamente elevados ou muito reduzidos, devido às consequências que podem gerar. Trata-se de um ramo da estatística que adquiriu maior relevância, principalmente, nos últimos setenta anos. Os seus domínios de aplicação são muito variados: meteorologia, seguros, telecomunicações, engenharia civil, economia, finanças, etc.

Esta Teoria também pode ser aplicada a acontecimentos mais correntes, como por exemplo, o fluxo de tráfego numa ponte.

É de conhecimento corrente que congestionamentos significativos, situações de "pára-arranca", provocam atrasos nas deslocações, desgastes nos veículos, aumentos de consumo de combustível e aumentos na duração das viagens. Todos estes fenómenos implicam custos significativos quer económicos quer de produtividade. Estes factos também se aplicam ao que acontece na Ponte 25 de Abril, em Lisboa, tendo um enormíssimo significado pela grande quantidade de veículos que diariamente a atravessam, por isso, foi considerado relevante efetuar este estudo, o qual está na base desta dissertação. Em concreto, é efetuada a aplicação da Teoria dos Valores Extremos ao tráfego da Ponte 25 de Abril, tal como indicado no título desta tese.

Esta dissertação está organizada do seguinte modo: no Capítulo 1 tem-se a introdução; o Capítulo 2 é constituído por três partes: na primeira é feita uma breve introdução à história da Ponte 25 de Abril e também de uma análise de todos os dados adquiridos referentes ao fluxo do tráfego dessa Ponte, na segunda, será analisada a sazonalidade nos dados, tendo presente que este não é o foco principal desta dissertação mas sim a aplicação da Teoria dos Valores Extremos ao estudo do tráfego na ponte, na terceira, será efetuada uma análise das evoluções dos preços das portagens entre os anos 1996 e 2019 e das receitas entre os anos 2003 e 2017, da Ponte 25 de Abril, disponibilizados; no Capítulo 3 está um resumo teórico de alguns modelos da Teoria dos Valores Extremos, tais como, o Modelo dos Valores Extremos Generalizado (GEV), o Modelo GEV Multivariado e o Modelo Generalizado de Pareto (GP); já no Capítulo 4 encontra-se o foco principal desta dissertação que é a aplicação da Teoria dos Valores Extremos aos dados do tráfego da Ponte 25 de Abril. Os dados mais trabalhados serão os dados diários do tráfego, desde 1 de janeiro de 2010 até 31 de dezembro de 2018. Por não serem os únicos valores que foram disponibilidade são também trabalhados os valores referentes ao tráfego médio mensal de 2006 a 2010 e os valores de tráfego médio anual de 1966 a 2006; no Capítulo 5 são efetuadas algumas observações finais sobre o estudo elaborado e indicados problemas em aberto que poderão ser estudados.

CAPÍTULO 2

Apresentação dos dados

2.1 Introdução

Esta secção tem três parte. Na primeira, apresenta-se um pouco da história da Ponte 25 de Abril e alguns acontecimentos (como por exemplo a construção da Ponte Vasco da Gama) que possam ter tido algum impacto na utilização da Ponte aqui estudada. Vão ser analisados todos os dados adquiridos e não só os referentes ao tráfego diário. Serão tidos em consideração os dados diários do tráfego da Ponte 25 de Abril desde 1 de janeiro de 2010 até 31 de dezembro de 2018, o tráfego médio diário mensal desde 2006 e o tráfego médio diário anual desde 1966.

Na segunda parte, será efetuada uma análise da existência ou não de sazonalidade nos dados. E na terceira e última parte, será elaborada uma análise referente ao valor pago por cada viatura dependendo da Classe que lhe é atribuída, como também, das receitas recebidas através do pagamento das mesmas.

2.2 Descrição geral dos dados

2.2.1 História da Ponte 25 de Abril

Nesta parte apresenta-se um pouco da história da Ponte 25 de Abril. Esta estrutura é uma ponte suspensa rodoferroviária que faz a ligação entre as cidades de Lisboa e Almada. Esta união é feita no denominado "gargalo do Tejo", isto é, na parte mais estreita e final do estuário do rio Tejo.

Na figura 2.1 está representado um cronograma com alguma da história da Ponte que teve como fonte de informação Infraestruturas de Portugal (2018a).

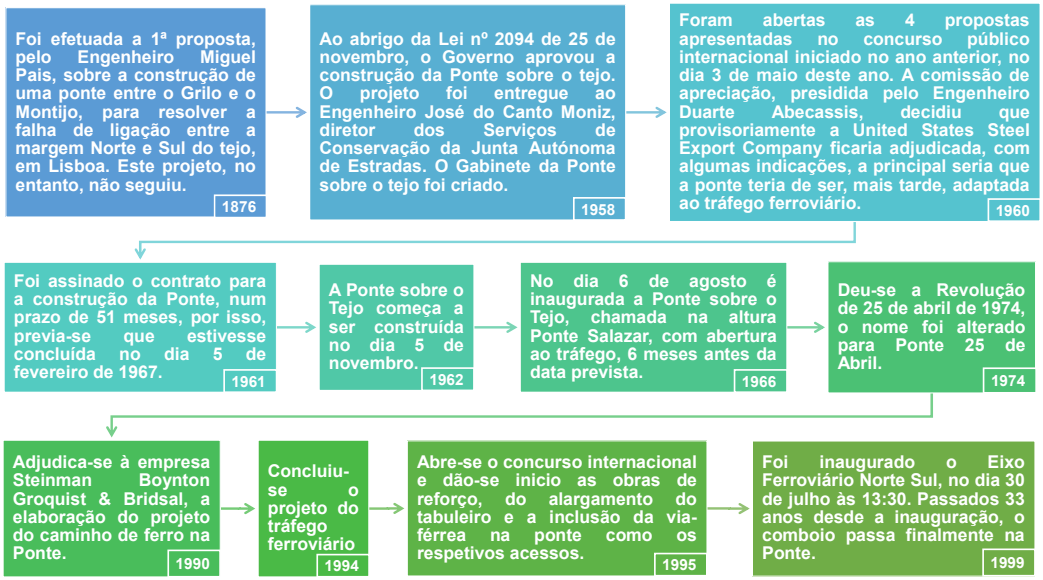


Figura 2.1: Cronograma da história da Ponte 25 de Abril resumida (1876-1999)

(a) Ponte 25 de Abril durante o dia (b) Ponte 25 de Abril durante a noite

Figura 2.2: Ponte 25 de Abril

Na figura 2.2 tem-se do lado esquerdo a figura 2.2a obtida através de Almeida (2018) e do lado direito a figura 2.2b obtida através de Costa (2018). Ao visualizar estas imagens pode-se ter uma noção mais adequada da grandeza da Ponte Salazar, como foi inicialmente chamada, se for ponderado o facto de que quando foi inaugurada, em 1966, era a quinta maior estrutura metálica do mundo e a maior fora dos Estados Unidos. Neste momento, ocupa o 36º lugar a nível mundial. Portanto, até esta altura só ainda construíram mais trinta e uma pontes maiores que a mencionada e já passaram cinquenta anos desde a sua inauguração. Por este andar, a Ponte sobre o Tejo ficará no Top50 até meados de 2040.

A Construção teve um custo que na altura rondou os dois milhões e duzentos mil contos, o que corresponde, ao valor aproximado de onze milhões de euros, sem ajuste de inflação (ou seja, sem ter em conta o aumento contínuo e generalizado dos preços).

Esta possibilidade de travessia, permitida pela Ponte 25 de Abril, teve como consequência uma explosão urbanística que surgiu na margem esquerda do Rio Tejo, de

Almada a Setúbal, e houve efetivamente um crescimento económico e turístico no Sul de Portugal, tendo como grande destaque a região do Algarve.

Como se pode observar no cronograma (figura 2.1) a travessia ferroviária foi inaugurada em 1999, mais concretamente no dia 30 de julho. E no ano anterior, no dia 29 de março, foi inaugurada a Ponte Vasco da Gama, uma nova travessia do Tejo, entre Sacavém e o Montijo. Estas duas modalidades diferentes de travessia, rodoviária e ferroviária, tinham como principal objetivo diminuir o tráfego da Ponte 25 de Abril, mas tal não ocorreu. Como se pode ler no trecho retirado do Volume I do documento "Auditoria à aplicação do Modelo Contratual e aos Acordos de Reposição do Equilíbrio Financeiro" efetuado pelo Tribunal de Contas Garcia, Pignatelli, Salina e Santos (2000):

"(...) a versão inicial do Modelo apresentava-se equilibrada tendo em atenção, entre outros, os seguintes pressupostos:

- Haveria uma diminuição do tráfego na Ponte 25 de Abril, na sequência da abertura da nova ponte, ou seja, a Ponte Vasco da Gama e da ferrovia.
- O tráfego previsto para 1998 e 1999, na Antiga Travessia, seria inferior ao estimado para 1996 e 1997.
- As taxas de portagem a cobrar na Ponte 25 de Abril, em 1998 e 1999, representariam valores superiores ao dobro, em termos médios, dos praticados em 1996 e 1997.

Contudo, a realidade mostrou que:

- Não houve uma diminuição do tráfego na Ponte 25 de Abril, na sequência da abertura da nova ponte e da ferrovia
- O tráfego na Ponte 25 de Abril, em 1998 e 1999, não foi inferior ao tráfego verificado em 1996 e 1997.
- As taxas de portagem cobradas em 1998 e 1999 representaram valores inferiores a metade dos constantes do Caso Base, isto é, mantiveram inalterado o seu valor. (...)"

A circulação tanto rodoviária como ferroviária é intensa. Todos os dias se ouvem nas notícias os congestionamentos recorrentes na Ponte sobre o Tejo. Dando particular destaque aos números, por exemplo conforme a notícia de Trainlogistic (2018): *"no início do ano 2006: na chamada "hora de ponta" passaram cerca de sete mil carros, nos dois sentidos e, em média, passavam por dia cerca de cento e cinquenta mil (...). Na mesma altura, em relação à circulação ferroviária, havia a passagem diariamente de cento e cinquenta e sete comboios, nos dois sentidos, transportando cerca de oitenta mil passageiros por dia. Em 2006 passavam cerca de quatrocentas mil pessoas na Ponte 25 de Abril."* Hoje em dia, só na parte rodoviária chegam a passar cerca de 140000 automóveis por dia nos dois sentidos.

2.2.2 Recolha e análise genérica do tráfego da Ponte 25 de Abril

Nesta dissertação, serão utilizados os dados referentes ao tráfego da Ponte 25 de Abril para a aplicação dos modelos da Teoria dos Valores Extremos. Depois de contactadas várias entidades como: a Brisa, a Lusoponte, o INE (Instituto Nacional de Estatística), e o IMT (Instituto da Mobilidade e dos Transportes, I.P.). A instituição que forneceu os dados apresentados e tratados neste trabalho foi o IMT a quem já se fez uma referência em particular nos agradecimentos. Pois sem estes, o estudo aqui apresentado não seria possível.

Obtiveram-se os valores diários referentes ao tráfego da Ponte 25 de Abril desde 1 de janeiro de 2010 até 31 de dezembro de 2018. O tráfego é contabilizado nos dois sentidos, apesar de só existirem portagens no sentido Sul-Norte, também é contabilizado o tráfego no outro sentido através de sensores colocados no pavimento. Foi a partir de 2010 que se começou a fazer a recolha dos dados diários pela própria entidade. Antes de 2010, só se têm em arquivo os dados das médias mensais até 2006 e, antes de 2006 até 1966, as médias anuais.

No próximo capítulo, os dados diários terão uma relevância particular, pois ao existirem variações ao longo do mês é necessário dispor dos dados desagregados para se aplicarem os modelos de valores extremos. Por isso, para não se deixarem dados de lado, será feita uma análise geral de todos os dados adquiridos.

Verifica-se a existência de 3287 valores diários; 156 valores referentes a TMDM (Tráfego Médio Diário Mensal), ou seja, a média mensal do tráfego diário; e 54 dados de TMDA (Tráfego Médio Diário Anual), isto é, a média anual do tráfego diário.

Na figura 2.3 pode-se observar a média anual do tráfego diário desde 1966 até 2018. Verifica-se que desde 1966 até 1993 têm um comportamento, tendencialmente, crescente. Têm um pequeno decréscimo em 1994 e outro em 1998. Este segundo poderá associar-se à inauguração da Ponte Vasco da Gama, no entanto, esta diminuição no tráfego não foi tão acentuada quanto se pretendia. Pois, como já se mencionou, a construção da segunda ponte para a travessia do Tejo era também para, de algum modo, reduzir o tráfego na Ponte 25 de Abril. Aliás, segundo a notícia do Jornal "Público" (2006), o trânsito aumentou 16% entre 1998 e 2005. No ano 1999 foi finalmente inaugurada a via férrea, mas, tal como se constata pela observação do gráfico, essa também não teve um impacto considerável no fluxo de veículos que atravessam a Ponte sobre o Tejo.

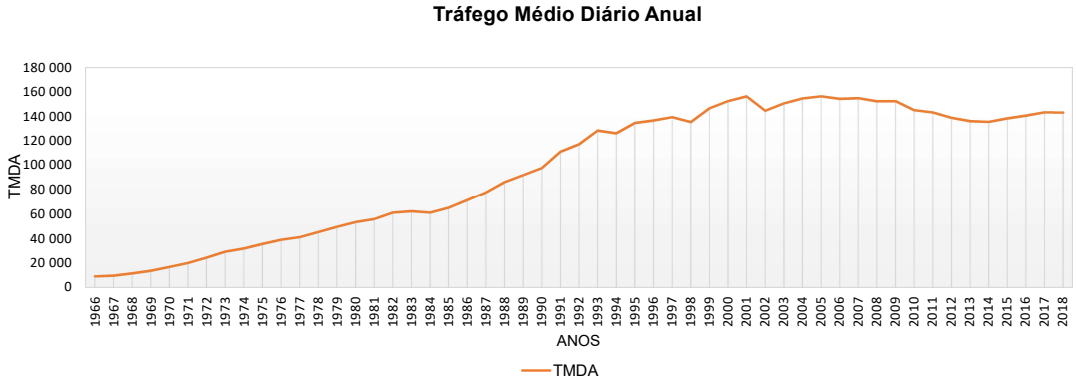


Figura 2.3: Tráfego Médio Diário Anual (1966-2018)

Através da figura 2.4 onde está representado o gráfico da média mensal do tráfego diário, desde 2006 a 2018, nota-se um ligeiro decréscimo no fluxo de automóveis de 2006 até 2014. Neste último ano, está o mínimo valor apresentado. Pode-se associar esta diminuição gradual ao impacto da Crise no poder de compra dos portugueses e que pode ter sido uma consequência imediata, a diminuição da utilização do automóvel.

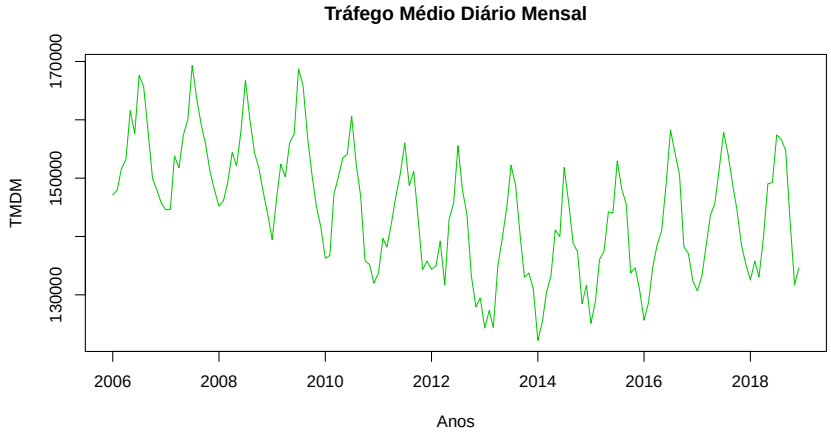


Figura 2.4: Tráfego Médio Diário Mensal (2006-2018)

No gráfico sequencial, apresentado na figura 2.5, estão representados os valores diários do tráfego da Ponte 25 de Abril. Nestes nota-se um comportamento que demonstra repetição na variação de valores. Logo pode-se dizer que os valores aparentam ter sazonalidade, esta será verificada no ponto seguinte desta secção.

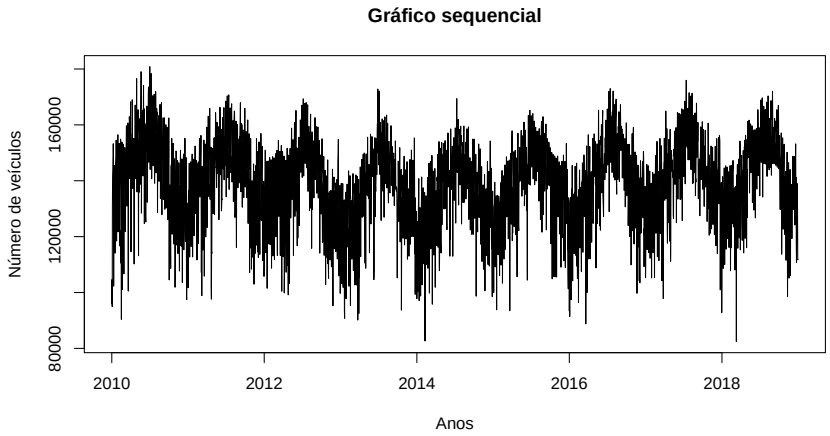


Figura 2.5: Gráfico sequencial de dados diários do tráfego da Ponte 25 de Abril (2010-2018)

2.3 Análise da sazonalidade

2.3.1 Apreciação Gráfica

Nesta parte procurar-se-á verificar a existência ou não de sazonalidade nos dados. Consegue-se através da figura 2.5, constatar a existência de uma certa oscilação que se pode considerar repetitiva. De modo mais detalhado, na figura 2.6, estão representados os dados diários de 2010.

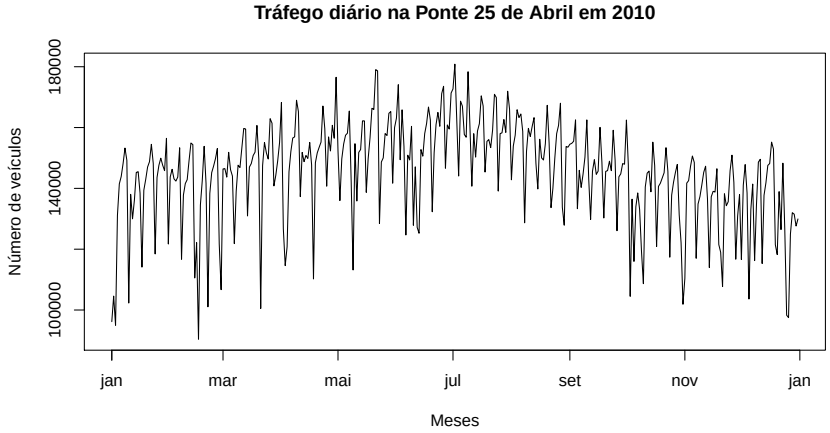


Figura 2.6: Tráfego diário na Ponte 25 de Abril em 2010

Pela figura 2.6 consegue-se visualizar que o fluxo de trânsito aumenta gradualmente até julho e mantém-se no seu momento máximo nesse mês e a partir de agosto, começa a diminuir. É bastante compreensível o comportamento do fluxo de tráfego, tendo em conta as estações do ano, por exemplo, consegue-se verificar que nos meses de inverno o tráfego na Ponte 25 de Abril é menor mas vai crescendo até aos meses de verão.

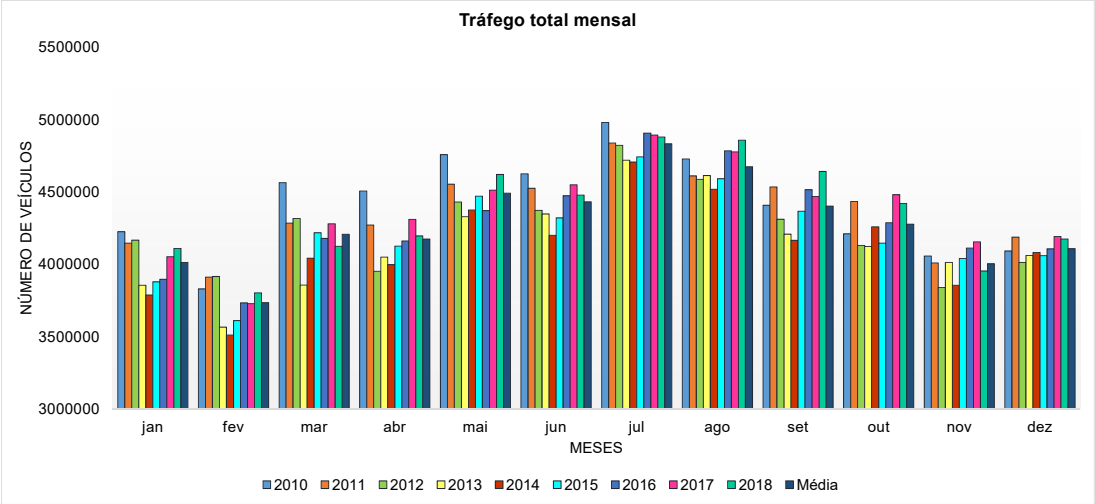


Figura 2.7: Tráfego total mensal na Ponte 25 de Abril (2010-2018)

Pela figura 2.7,verifica-se a variação existente em cada mês, tendo no gráfico do fluxo total mensal de cada ano (de 2010 a 2018) e a média mensal. O número de veículos a fazer a travessia do Tejo tem o seguinte comportamento: vai aumentando gradualmente de janeiro a julho onde há um "pico de tráfego" notório; a partir daqui há uma diminuição de agosto a novembro; por fim, um pequeno aumento em Dezembro. Tal como se constatou na figura 2.6 mas não com tanto detalhe.

Ano/mês	JAN	FEV	MAR	ABR	MAI	JUN	JUL	AGO	SET	OUT	NOV	DEZ
2010	4224652	3828746	4563546	4506398	4757528	4624064	4980058	4727474	4406776	4209592	4056474	4091686
2011	4145533	3910751	4284042	4270360	4553549	4524434	4837444	4610856	4533814	4433562	4008582	4187329
2012	4166050	3914763	4315830	3950928	4430800	4372117	4821938	4586311	4311110	4129951	3838369	4012693
2013	3854645	3565231	3855582	4049453	4328235	4347041	4718802	4612927	4207995	4123280	4012741	4060121
2014	3787677	3510724	4041857	3997244	4374994	4199051	4706455	4518544	4164528	4258606	3854476	4080572
2015	3878744	3609397	4217205	4124907	4470700	4320618	4741665	4590546	4366085	4146466	4038909	4059142
2016	3895119	3732524	4178028	4160492	4370279	4473275	4906518	4783522	4515494	4286352	4111362	4105737
2017	4051244	3727218	4279434	4309458	4512357	4549332	4893114	4776071	4467813	4480308	4154925	4191607
2018	4108993	3801563	4124483	4195723	4619924	4477089	4879373	4857197	4641572	4420293	3953403	4173989
Média	4012517	3733435	4206667	4173885	4490930	4431891	4831707	4673716	4401687	4276490	4003249	4106986

Tabela 2.1: Valores do tráfego total mensal na Ponte 25 de Abril (2010-2018) e Média mensal

Na Tabela 2.1 tem-se a amarelo o valor anual mais pequeno, pode-se dizer que normalmente o fluxo de trânsito é mais pequeno em meses com temperaturas mais baixas, já que estes valores ocorreram sempre em fevereiro (se bem que este mês como tem menos dias é normal que o seu valor total mensal seja mais reduzido), excepto em 2012 que foi em novembro. A vermelho está representado o maior valor anual, que tem sido sempre no mês de julho. A verde está representado o maior valor mensal, em cinco dos doze meses ocorreram em 2010, um dos valores ocorreu em 2012, três em 2017 e dois em 2018. Já os menores valores mensais, representados a azul, oscilaram entre os anos de 2012 a 2014,

sendo que em 2014 teve seis meses com os menores valores. Os valores a cor-de-laranja são duplamente representativos: em fevereiro de 2014 é o menor valor mensal e anual; em julho do mesmo ano está representado o maior valor anual e o menor valor mensal.

Máximos anuais			
Ano	**Data**	**Dia da semana**	**Número veículos**
2010	02/jul	sexta-feira	180 846
2011	15/jul	sexta-feira	170 750
2012	06/jul	sexta-feira	169 322
2013	28/jun	sexta-feira	172 842
2014	11/jul	sexta-feira	169 406
2015	26/jun	sexta-feira	165 212
2016	15/jul	sexta-feira	172 982
2017	14/jul	sexta-feira	175 961
2018	31/ago	sexta-feira	172 030

Tabela 2.2: Datas e dias da semana dos valores máximos anuais

Quanto à tabela 2.2 pode-se afirmar que os valores máximos anuais ocorrem sempre entre a última semana de junho e a primeira quinzena de julho, excepto em 2018, que foi no último dia do mês de agosto, e todos estes valores máximos anuais ocorreram numa sexta-feira.

2.3.2 Ajuste Sazonal

A sazonalidade é recorrentemente causada por movimentos que possuem a mesma periodicidade e, normalmente, oscilatórios e ocorrem em períodos determinados no meio do ano, como feriados, variações climáticas, férias, etc. O Processo de remoção da sazonalidade de uma série temporal é conhecido como ajuste sazonal. Nesta secção será apresentado, tendo como base o artigo Ferreira e Mattos (2016) que usa o X-13ARIMA-SEATS com interface no *software R* utilizando o pacote *seasonal* (Sax & Eddelbuettel, 2018) que foi desenvolvido por Christoph Sax, um modo de dessazonalizar séries temporais.

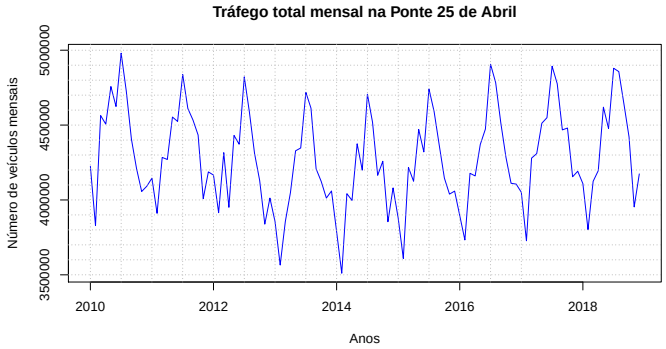


Figura 2.8: Tráfego total mensal na Ponte 25 de Abril

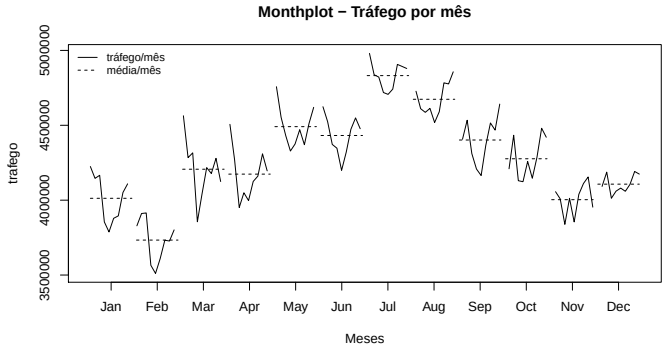


Figura 2.9: *Monthplot* - Tráfego por mês na Ponte 25 de Abril

Ao analisar as figuras 2.8 e 2.9 pode-se verificar que o índice de fluxo de tráfego:

a. Tem características sazonais, já que de fevereiro a julho tem um comportamento maioritariamente crescente e de julho a fevereiro, decrescente. E este comportamento verifica-se ao longo dos anos;

b. Houve um pico mínimo bastante acentuado em 2013, talvez nessa altura se tenha sentido de forma mais acentuada as consequências da crise económica em Portugal. Se bem que em 2014 existe um mínimo menor que o de 2013 como se pode observar no gráfico 2.8;

c. Através do gráfico 2.8, verifica-se que existe uma tendência decrescente de 2010 a 2014 e de 2014, até ao fim deste gráfico, a tendência passa a ser crescente, talvez a partir desta altura as consequências da crise não se sintam de modo muito acentuado;

d. Pelo gráfico 2.9 com o título "Monthplot -Tráfego por mês" (já que *Monthplot* significa o gráfico mensal), em que se tem a série temporal de cada mês do ano (ou seja, estão juntos os dados dos anos de 2010 a 2018, por mês) verifica-se que os dois maiores valores mensais ocorrem nos meses de julho e agosto, que acabam por ficar

justificados pela altura do ano, já que no verão, na altura balnear, a passagem pela Ponte 25 de Abril é mais recorrente dado o número de praias existentes na margem Sul.

Depois de analisado o comportamento histórico da Série Temporal, efetuou-se o ajuste sazonal automático, usando o X13-ARIMA-SEATS que a partir de agora será denominado por X13. Como existem vários estudos empíricos que mostram que nem todos os ajustes automáticos conseguem remover a sazonalidade, como seria esperado, é muito importante fazer o teste de sazonalidade. No X13 é dado pela estatística QS. O teste tem como hipótese nula: não existe sazonalidade. Para explicar melhor esta estatística foram tidas em conta as informações de Bureau (2017), "StackExchange" (2018) e Maravall (2005) e resumindo calcula-se do seguinte modo:

1. A série para a qual é calculada a estatística QS é diferenciada de acordo com o modelo ARIMA (que será explicado com maior detalhe mais adiante) e também pela seguinte regra:

$$ndif = max(1, min(d + D, 2))$$

onde:

$ndif$: é o número de diferenças regulares a serem tidas em conta;

d e D: são, respetivamente, o número de diferenças regulares e sazonais no modelo ARIMA escolhido.

(O $ndif = 0$ irá ocorrer se a estatística QS for calculada para a série de resíduos, ou seja, nenhuma diferença será aplicada.)

2. As duas primeiras autocorrelações de ordem sazonal (em dados mensais, como neste caso, serão 12 e 24) são obtidas e se essas autocorrelações forem menores ou iguais a zero, então serão definidas como zero.

3. A estatística é definida do seguinte modo:

$$QS = n(n+2)\left(\frac{R_s^2}{n-2} + \frac{R_{2s}^2}{n-2s}\right)$$

onde:

n: número de observações das séries diferenciadas;

s: é a periodicidade dos dados (12, neste caso, com os dados ordenados mensalmente);

R_s^2 e R_{2s}^2: são as autocorrelações obtidas no ponto anterior.

Apresenta-se para a estatística QS calculada para a série original, tendo em conta que os dados foram agrupados mensalmente, o *código do R* que se encontra em anexo I.1.1.

O *output* do teste da estatística QS que está apresentado na tabela 2.3 mostra que o teste para além de ter sido efetuado à série original e com ajuste, também foi aplicado nas séries de resíduos do modelo ARIMA e da componente irregular. Tem-se a expectativa de que não existam evidências de sazonalidade em todas as séries, exceto na série original. Como se pode observar dado o grande valor da estatística QS (e o baixo valor do p-value implícito) pode-se concluir que há sazonalidade na série.

	qs	valor p
Série original	117,9867	0,0000
Série original corrigida por valores extremos	124,4607	0,0000
Série dos resíduos do modelo ARIMA	0,0000	1,0000
Série temporal com ajuste sazonal	0,0000	1,0000
Série temporal com ajuste sazonal corrigida por valores extremos	0,0000	1,0000
Série de componente irregular	0,0000	1,0000
Série de componente irregular corrigida por valores extremos	0,0000	1,0000
Série original	102,5000	0,0000
Série original corrigida por valores extremos	110,3575	0,0000
Série dos resíduos do modelo ARIMA	0,0000	1,0000
Série temporal com ajuste sazonal	0,0000	1,0000
Série temporal com ajuste sazonal corrigida por valores extremos	0,0000	1,0000
Série de componente irregular	0,0000	1,0000
Série de componente irregular corrigida por valores extremos	0,0000	1,0000

Tabela 2.3: *Output* da Estatística QS

O teste é feito na série completa (resultados da linha 2 à 8 do *output*) e nos últimos 8 anos mais recentes (já que neste caso o comprimento é maior que 8 anos e está apresentado nas linhas seguintes do *output*), caso contrário, o teste seria feito apenas na série completa.

Ao considerar um nível de confiança de 95%, não existe nenhuma evidência de sazonalidade nas séries dessazonalizada, dos resíduos do modelo ARIMA e da componente irregular. No entanto, existem evidências de sazonalidade para a série original. De seguida diagnostica-se o pré-ajuste e o modelo ARIMA. Mas antes disso, uma breve explicação sobre o modelo ARIMA, para tal teve-se em consideração a explicação dada em Wikipedia (2018).

ARIMA é uma sigla em inglês para *"autoregressive integrated moving average"*, ou seja, para um modelo autorregressivo integrado de médias móveis. Este modelo é uma generalização do modelo ARMA (modelo autorregressivo de médias móveis). Estes dois modelos são ajustados aos dados da série temporal para compreender melhor o comportamento dos dados ou para efetuar previsões de futuros pontos na série. Os modelos ARIMA são aplicados normalmente a dados que mostram a não evidência de estacionariedade, por isso, a série dos dados inicial é diferenciada uma ou mais vezes para eliminar a não estacionariedade.

A parte autorregressiva (AR) do modelo ARIMA indica que a variável de interesse (que evolui) é regredida nos seus próprios valores desfasados (ou seja, anteriores). A parte integrada (I) indica que os valores dos dados foram substituídos pela diferença entre valores (ou seja, por exemplo, os valores de X_2 serão substituídos pela diferença entre X_2 e X_1) e este processo diferenciador pode ter sido realizado mais do que uma vez. A parte da média móvel (MA) indica que o erro de regressão é na verdade uma combinação linear dos termos de erro, cujos valores ocorreram simultaneamente, e em vários momentos no passado. O objetivo de cada uma destas características é elaborar um modelo que se ajuste aos dados do melhor modo possível.

Os Modelos ARIMA não sazonais são geralmente denotados por $ARIMA(p,d,q)$, em que os parâmetros **p**, **d** e **q** são números inteiros não negativos, **p** é a ordem (número de desfasamentos) do modelo autorregressivo, **d** é o grau de diferenciação (número de vezes em que os dados tiveram valores passados, subtraídos) e **q** é a ordem do modelo da média móvel. Os Modelos ARIMA sazonais são denotados normalmente como $ARIMA(p,d,q)(P,D,Q)_m$, em que m se refere ao número de períodos de cada intervalo e **P**, **D** e **Q** referem-se aos termos de autorregressão, diferenciação e média móvel para a parte sazonal do modelo ARIMA, respetivamente.

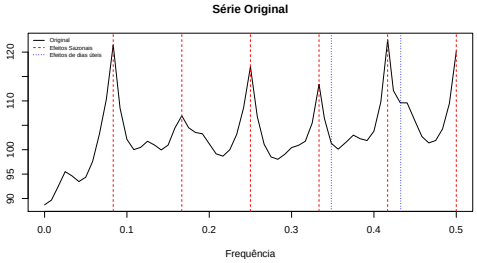
Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	Codes
Constant	2054	1124	1.827	0.0677	0.05
Easter[1]	-82760	39020	-2.121	0.0339	0.01
MA-Nonseasonal-01	0,6925	0,06691	10.349	<2e-16	0
MA-Seasonal-12	0,9978	0,077974	12.513	<2e-16	0
SEATS adj. ARIMA:	(0 1 1)(0 1 1)		Obs.:	108	
Transform:	none	QS (no seasonality in final):		0	
AICc:	2477	Box-Ljung (no autocorr.):		35.87	0.05
BIC:	2489	Shapiro (normality):		0.9797	0.05

Tabela 2.4: *Output* do *summary(ajuste)*

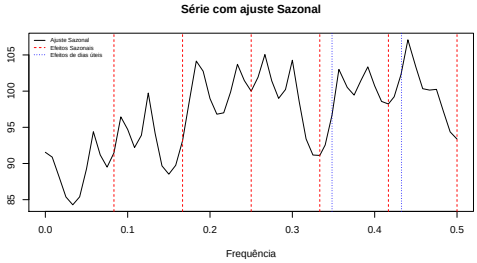
Em relação ao *output* obtido pode-se afirmar o seguinte: o modelo ARIMA estimado é da ordem (0 1 1)(0 1 1) e o parâmetro MA sazonal é significativo. De acordo com o teste de autocorrelação de Box-Ljung, não existem evidências de autocorrelação residual para o modelo ARIMA estimado. O teste de normalidade de Shapiro-Wilk sugere a não existência de normalidade, no entanto, essa não é uma característica extremamente necessária no diagnóstico de modelos ARIMA. Também se verifica que não foi empregue qualquer transformação logarítmica.

O próximo passo é fazer um diagnóstico, fornecido pelo programa, que tem o objetivo de verificar se existem indícios de sazonalidade e efeitos de dias úteis antes e depois do ajuste sazonal. Este diagnóstico é fornecido pelo gráfico da função de autocovariância, de uma dada Série Temporal, reestimada por densidade espectral. Este diagnóstico é dado

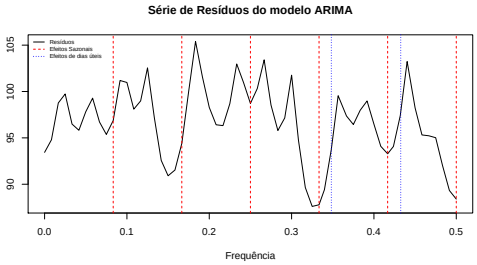
por quatro séries que são: a série original, a série dessazonalizada, a série dos resíduos do modelo ARIMA e a série da componente irregular. No *R*, usa-se a função *series()* do pacote *seasonal* para obter as séries espectrais.



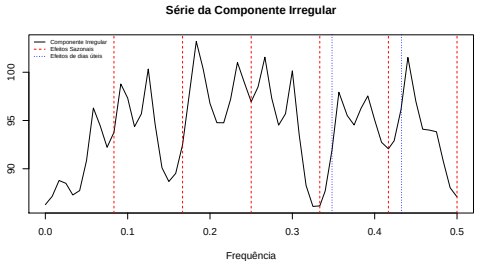
(a) Série Original



(b) Série com ajuste Sazonal



(c) Série de Resíduos do modelo ARIMA



(d) Série da componente Irregular

Figura 2.10: Gráficos espectrais para efeitos de sazonalidade e dias úteis

Para serem interpretados os gráficos representados na figura 2.10 tem que se ter em conta o seguinte: existem indícios de efeitos sazonais na série, se a densidade espectral da Série Temporal apresenta mais do que um pico sobre frequências sazonais (linhas vermelhas e tracejadas); também existem indícios de efeitos de dias úteis, caso hajam picos nas frequências de dias úteis (linhas tracejadas em azul). Como tal, pode-se afirmar que só se verificam efeitos de sazonalidade para a série original e não existem efeitos de dias úteis. Por esse motivo, não é necessário corrigir nenhum efeito referente aos dias úteis. E verifica-se que a série está bem dessazonalizada, já que só o primeiro gráfico apresenta mais do que um pico sobre frequências sazonais.

Os dois gráficos apresentados de seguida, nas figuras 2.11 e 2.12, são referentes aos factores sazonais e à série com ajuste sazonal.

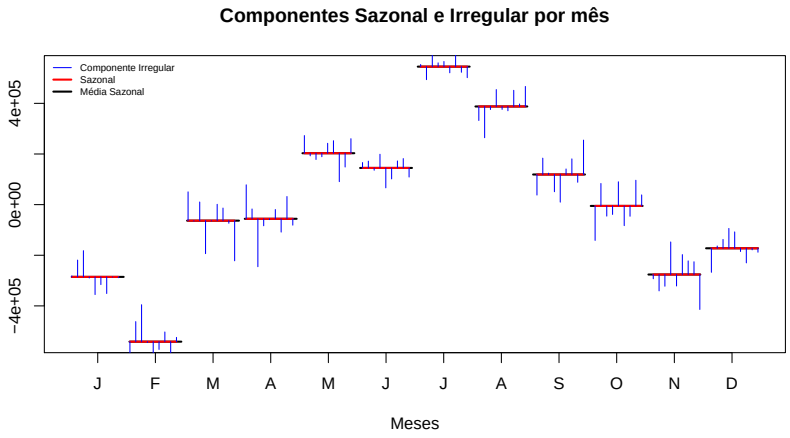


Figura 2.11: Componentes Sazonal e Irregular por mês

A representação gráfica da figura 2.11 é útil para visualizar a evolução dos fatores sazonais ao longo do tempo e é dada pela função *monthplot()*. Para além da evolução dos fatores sazonais, é possível verificar o comportamento da série SI (componentes sazonal e irregular agregadas).

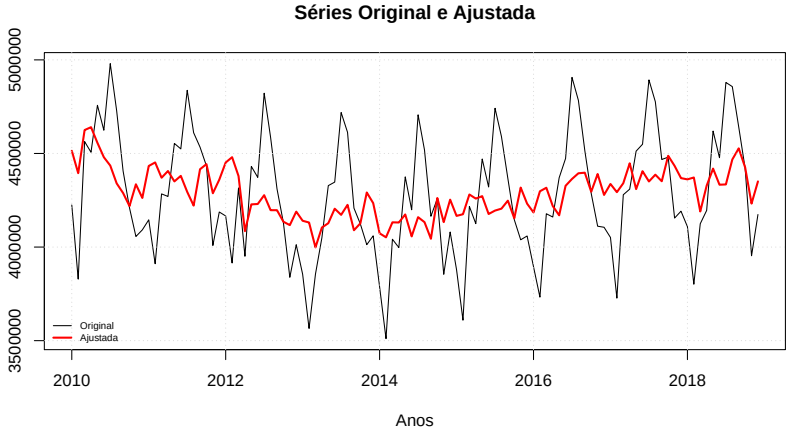


Figura 2.12: Séries Original e Ajustada

Na figura 2.12 vê-se o fluxo de tráfego na Ponte 25 de Abril com ajuste sazonal, através do qual, mais uma vez, se constata uma ligeira descida até 2014 do número de automóveis a atravessar a Ponte e um aumento gradual nos anos seguintes.

O programa X13, usando modelos SARIMA (ARIMA Sazonal, ou seja, em inglês *Seasonal ARIMA*) faz previsões não só da Série original mas também da Série com ajuste sazonal. Do adequado ajuste sazonal depende a qualidade das previsões efetuadas a partir deste.

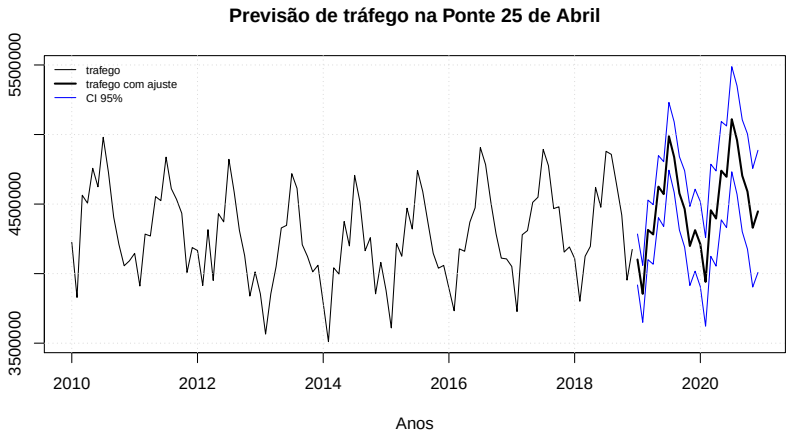


Figura 2.13: Primeira previsão do tráfego na Ponte 25 de Abril com ajuste sazonal

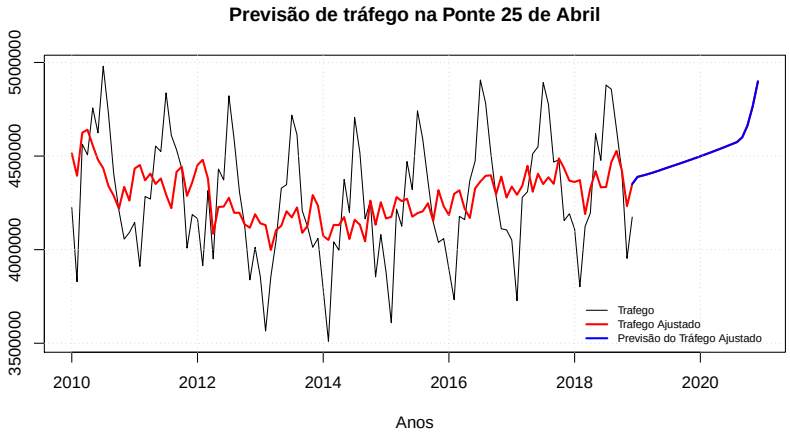


Figura 2.14: Segunda previsão do tráfego na Ponte 25 de Abril com ajuste sazonal

Nas figuras 2.13 e 2.14 encontram-se representadas as previsões do tráfego com ajuste sazonal. Na figura 2.13, tem-se a preto o tráfego com ajuste representado e a azul os respetivos IC de 95% dessa previsão. Verifica-se que a previsão indica que o máximo de 2020 será superior aos máximos anuais dos anos anteriores. Pela visualização do gráfico 2.14 é notória a tendência crescente da previsão do tráfego ajustado. Os resultados previstos estão de um modo mais detalhado nas tabelas I.1 e I.2 em anexo.

2.4 Análise do valor e receitas das portagens da Ponte 25 de Abril

Nesta secção, vai-se procurar fazer uma pequena análise das receitas obtidas na passagem da Ponte 25 de Abril, algumas das informações apresentadas têm como origem os seguintes documentos: Infraestruturas de Portugal (2017) e Infraestruturas de Portugal

(2018b). Inicialmente ter-se-á em conta o valor unitário das portagens e a sua evolução. De seguida, vai se fazer uma breve análise das receitas realizadas.

No decorrer do texto seguinte será utilizado o conceito de "valor unitário", entendido como o valor pago por cada viatura dependendo da Classe que lhe é atribuída.

Antes da perspectiva mais detalhada sobre o valor unitário das portagens, vão-se esclarecer as seguintes observações:

1. A distinção entre cada uma das Classes de veículos é apresentada na tabela 2.5 que foi retirada do site Lusoponte (2019a).

CLASSE	ALTURA VERTICAL AO 1º EIXO	Nº DE EIXOS	TIPO DE VEÍCULO
1	< 1,1m	=> 2	V eículos com uma altura, medida à vertical do primeiro eixo, inferior a 1,10m, com ou sem reboque
2	=> 1,1m	2	Veículos com dois eixos e uma altura, medida à vertical do primeiro eixo, igual ou superior a 1,10 m
3	=> 1,1m	3	Veículos com três eixos e uma altura, medida à vertical do primeiro eixo, igual ou superior a 1,10 m
4	=> 1,1m	=> 4	Veículos com mais de três eixos e uma altura, medida à vertical do primeiro eixo, igual ou superior a 1,10 m

Tabela 2.5: Descrição dos veículos de cada uma das Classes

   Atualmente, consideram-se da Classe 5 os motociclos, que pagam de portagem o mesmo que a Classe 1. Mas para esta análise aqui apresentada não serão tidos em consideração por falta de informações.

2. Houve uma mudança na zona de cobrança da Ponte 25 de Abril do sentido Norte-Sul para o sentido Sul-Norte na madrugada do dia 28 de novembro de 1992. E as portagens sempre foram cobradas só num dos sentidos.

3. Aqui serão apresentados os dados, do valor unitário das portagens, a partir de 1992, pois foram os dados fornecidos pelo IMT.

4. Até 2010 inclusive as portagens não eram cobradas no mês de agosto na Ponte 25 de Abril, por ser o período de férias dos operadores desta portagem. Mas a partir de 2011 passou a ser cobrada e assim, em princípio, se irá manter.

Passando à análise do valor unitário das portagens pagas por cada uma das quatro Classes de veículos têm-se os dados referentes nos gráficos 2.15 e 2.16. Os dados de 1996

a 2001 foram convertidos para euros, já que os seus valores originais estavam em escudos.

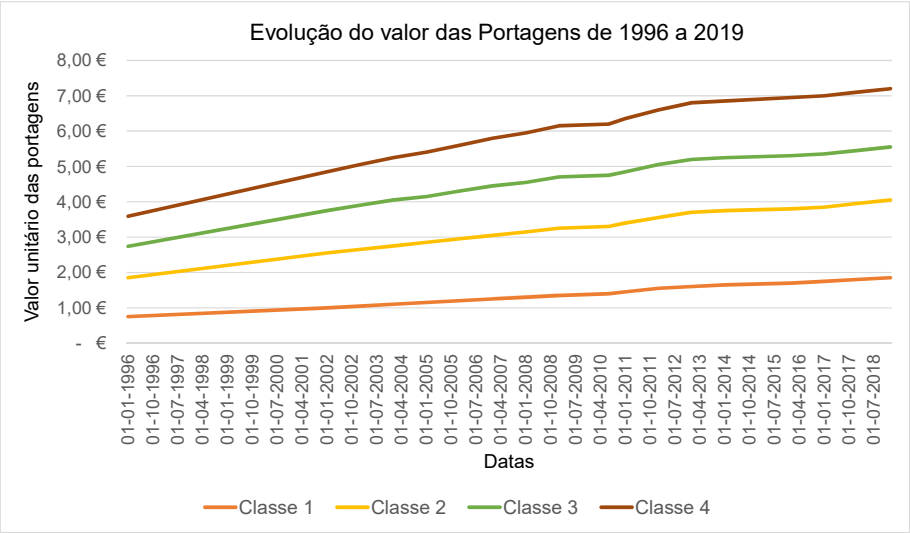


Figura 2.15: Evolução do valor das Portagens de 1996 a 2019 da Ponte 25 de Abril

Ao se observar o gráfico 2.15 constata-se que o valor unitário das portagens tem vindo a aumentar ao longo dos anos. Este aumento, aparenta ser praticamente constante. Só em 2010 se nota uma pequena diminuição no aumento, ou seja, a diferença entre o valor das portagens de um ano para o outro foi menor. Por exemplo, de 2008 para 2009 houve um aumento de vinte cêntimos, no valor unitário da portagem, para a Classe 4; de 2009 para 2010 não houve nenhuma diferença no início do ano (como nos restantes anos tinha havido) e, mais tarde, no segundo semestre de 2010, houve um aumento de apenas cinco cêntimos para a Classe 4.

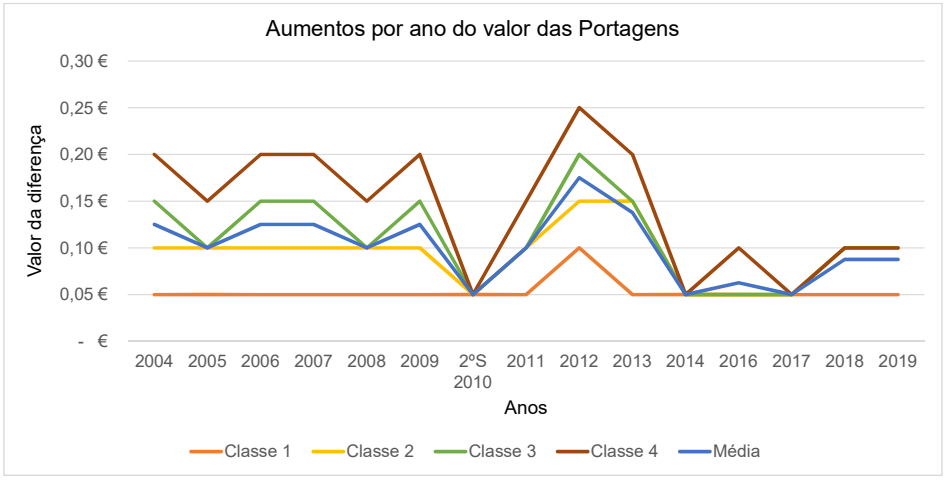


Figura 2.16: Aumentos por ano do valor unitário das Portagens da Ponte 25 de Abril

Em relação ao gráfico 2.16 que se refere às diferenças entre o valor das portagens dos vários anos, como estas têm sido sempre positivas, o gráfico denomina-se como "aumentos por ano do valor das portagens". Este gráfico apresenta dois anos relevantes. O primeiro já foi visualizado no gráfico 2.15, em relação ao segundo semestre de 2010, houve uma grande diminuição no valor do aumento, já que as quatro Classes só aumentaram cada uma cinco cêntimos. O segundo ano mais relevante é 2012, onde se nota um aumento bastante acentuado no valor unitário das Portagens. Por exemplo, a Classe 4 aumentou nesse ano vinte e cinco cêntimos. Os dados estão apresentados de modo mais detalhado em tabelas que se encontram em anexo, em relação ao gráfico 2.15 na tabela I.1 e na tabela I.2 sobre os dados do gráfico 2.16.

Quanto às receitas da Ponte 25 de Abril obtidas podem-se observar os gráficos 2.17 e 2.18 cujos valores representados são os valores reais cobrados e foram adquiridos através do "INE" (2018). De um modo mais detalhado as receitas estão representadas em duas tabelas em anexo na I.3 e na I.4.

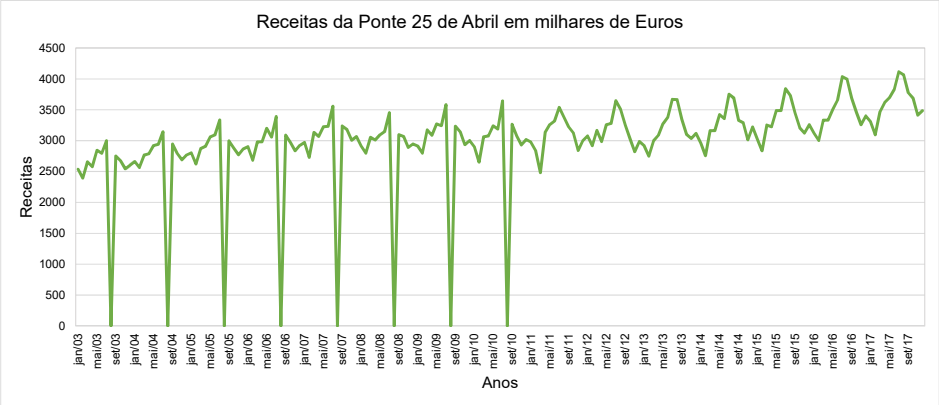


Figura 2.17: Receitas totais mensais da Ponte 25 de Abril (2003-2017)

É visível através da figura 2.17 que as receitas obtidas através da Ponte 25 de Abril têm uma tendência crescente. Em concreto, nesta representação, nota-se o impacto bastante acentuado no mês de agosto (neste caso, nos anos 2003 a 2010) já que esta portagem não era cobrada nesse mês, como já foi referido. A 30 de julho de 2012 saiu a seguinte notícia no "Jornal de Negócios" (2012), referente a esta mudança:

"A isenção de portagens em Agosto começou em 1996 e resultou da renegociação do contrato de concessão entre o Estado e a Lusoponte depois do bloqueio na Ponte sobre o Tejo, a que se chamou "buzinão".

Segundo o secretário de Estado das Obras Públicas, Transportes e Comunicações, Sérgio Monteiro, as isenções na cobrança de portagens nos meses de agosto desde essa altura geraram uma dívida de 110 milhões de euros.

No ano passado, o Governo decidiu reintroduzir as portagens devido às "dificuldades financeiras que o país atravessa" e aos "compromissos de redução de despesa pública assumidos pelo Estado português".

Esta medida vai vigorar todos os anos até ao termo da concessão da ponte, em 2030, segundo fonte da Lusoponte.(...)"

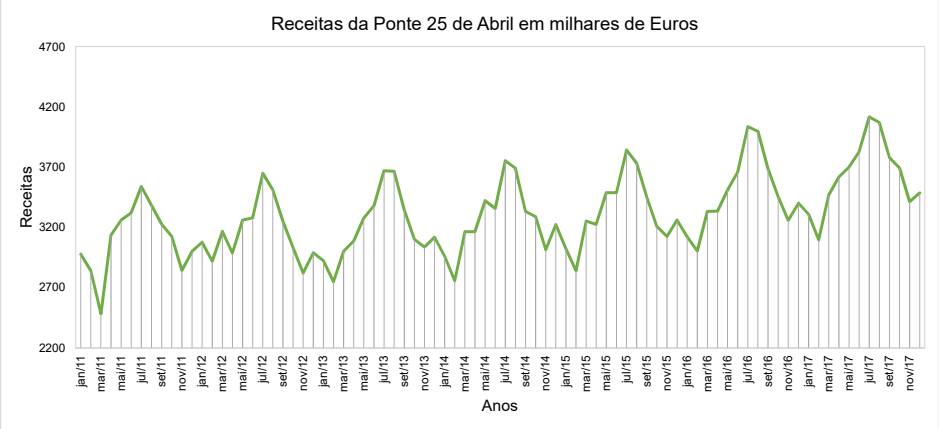


Figura 2.18: Receitas totais mensais da Ponte 25 de Abril, de 2011 a 2017

O gráfico 2.18 apresenta com maior detalhe o comportamento das Receitas da Ponte 25 de Abril de 2011 a 2017. Tal como os dados referentes ao fluxo de veículos na Ponte 25 de Abril as receitas, naturalmente, mostram o mesmo comportamento. Por isso, é visível um comportamento semelhante em todos os anos, ou seja, há um aumento acentuado de fevereiro até julho, onde há um pico de receitas, e depois decresce até novembro, tem um pequeno aumento em dezembro que se pode associar às festividades deste mês (como o Natal e a passagem de ano) e volta a diminuir até fevereiro. Não deixa de ser notório o fluxo elevado de receitas provenientes da Ponte 25 de Abril onde, por exemplo, no mês de julho de 2017 foram de 4116 milhares de Euros.

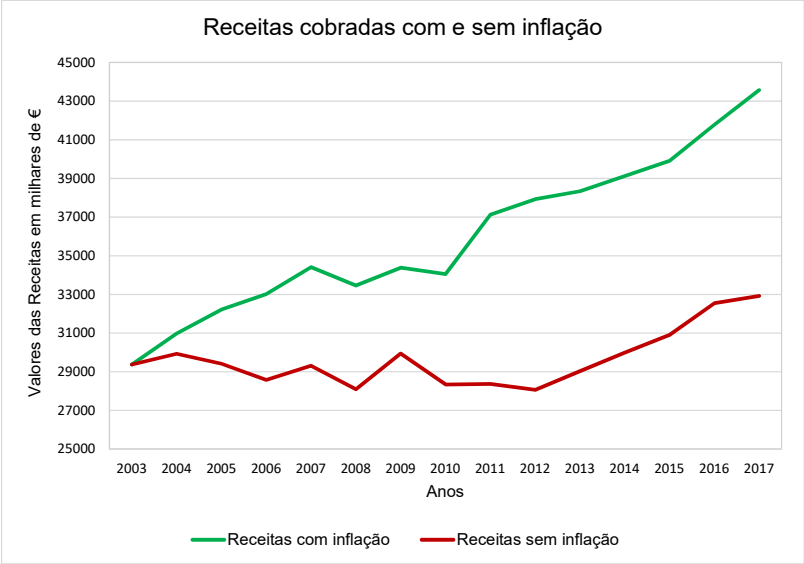


Figura 2.19: Receitas totais anuais cobradas na Ponte 25 de Abril, com e sem inflação a preços constantes de 2003 (2003-2017)

Se se analisar, anualmente, os valores têm vindo a aumentar como se pode observar na figura 2.19 mesmo que se retire a inflação ao valor das receitas cobradas (para se ver com maior detalhe cada um dos valores pode-se observar em anexo a tabela I.5). Já na figura 2.20 verifica-se a diferença entre as receitas anuais.

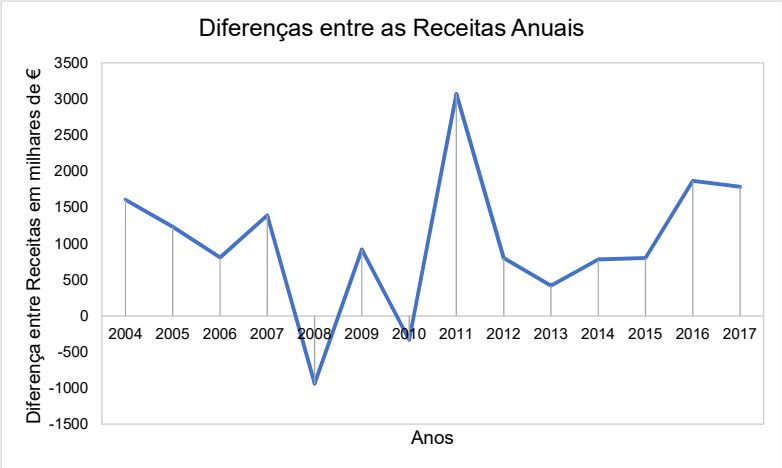


Figura 2.20: Diferenças das receitas totais anuais cobradas da Ponte 25 de Abril (2003-2017)

As diferenças entre estes anos são quase sempre positivas, ou seja, houve quase sempre aumentos nas receitas totais cobradas na Ponte 25 de Abril, com excepção da diferença entre os anos 2007-2008 e de 2009-2010, já que no primeiro caso houve uma diminuição de 940 milhares de euros e no segundo uma diminuição de 332 milhares de euros. Podem observar-se, com maior detalhe, as diferenças entre as receitas anuais na tabela I.6 apresentada em anexo.

Meses / Tráfego/receita	Unidade	Anual	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez
Tráfego médio diário (a)	nº	205 881	186 144	189 674	197 827	204 834	208 285	215 334	226 256	221 052	215 614	208 052	200 586	195 765
Ponte 25 de Abril	"	143 542	130 685	133 115	138 046	143 649	145 560	151 644	157 842	154 067	148 927	144 526	138 498	135 213
Ponte Vasco da Gama	"	62 339	55 459	56 559	59 781	61 185	62 725	63 690	68 414	66 985	66 687	63 526	62 088	60 552
Receita cobrada	10^3 Eur	78 825	6 006	5 534	6 351	6 424	6 739	6 814	7 373	7 263	6 888	6 761	6 324	6 348
Ponte 25 de Abril	"	43 569	3 308	3 095	3 469	3 618	3 699	3 828	4 116	4 069	3 778	3 691	3 413	3 485
Ponte Vasco da Gama	"	35 256	2 698	2 439	2 882	2 806	3 040	2 986	3 257	3 194	3 110	3 070	2 911	2 863

(a) Soma do tráfego médio diário realizado em cada uma das pontes.

Fonte: Instituto da Mobilidade e dos Transportes

Tabela 2.6: Tráfego médio diário e receitas cobradas nas pontes 25 de Abril e Vasco da Gama, de janeiro a dezembro de 2017 e a soma anual

A tabela 2.6 foi retira do relatório Lima (2018). Nesta tabela consegue-se observar, em relação às receitas, o valor total cobrado em cada uma das pontes 25 de Abril e Vasco da Gama, como também, a soma das mesmas. Nota-se que a Ponte 25 de Abril acaba por ter sempre um maior volume de receitas e para uma visualização mais detalhada tem-se o gráfico 2.21.

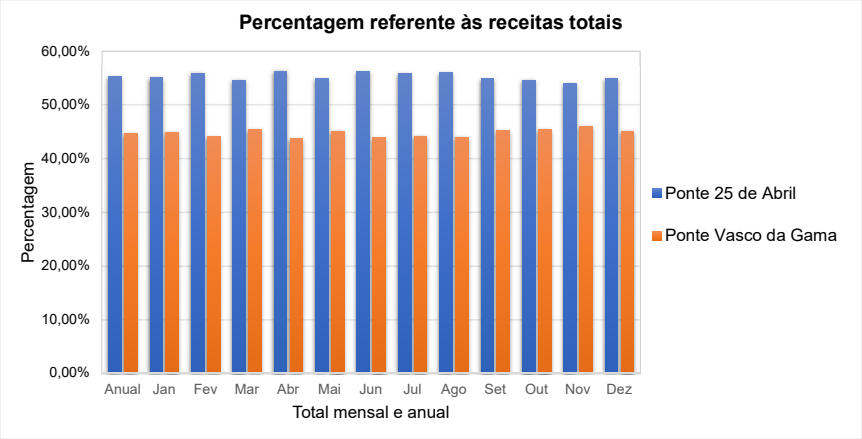


Figura 2.21: Percentagem referente às receitas totais da Lusoponte em 2017

No gráfico 2.21 é interessante verificar que a maioria das receitas totais da Lusoponte são provenientes da Ponte 25 de Abril. Também se nota, através da tabela 2.6, que existe mais fluxo de veículos a atravessar a Ponte 25 de Abril do que a Ponte Vasco da Gama. É possível que o valor unitário da Portagem que se paga tenha alguma relevância na decisão dos utilizadores sobre qual a ponte que irão usar para chegarem ao seu destino. Em 2019, na Ponte Vasco da Gama os preços das portagens são os apresentados na tabela 2.7, como se pode verificar em Lusoponte (2019b).

	Ponte 25 de Abril	Ponte Vasco da Gama	Diferença entre valores
Classe 1	1,85 €	2,85 €	1,00 €
Classe 2	4,05 €	6,45 €	2,40 €
Classe 3	5,55 €	9,50 €	3,95 €
Classe 4	7,20 €	12,20 €	5,00 €
Média	4,66 €	7,75 €	3,09 €

Tabela 2.7: Portagens pagas em cada uma das pontes da Lusoponte e respetivas médias (valores de 2019)

Concluindo, em 2019, em média, na Ponte Vasco da Gama paga-se mais 3,09 euros que na travessia da Ponte 25 de Abril. No entanto, não se irá entrar em grandes detalhes já que este não é o ponto fundamental desta dissertação.

CAPÍTULO 3

A Teoria dos Valores Extremos

3.1 Introdução

O estudo que nesta tese vai ser apresentado utiliza, como instrumento, a Teoria dos Valores Extremos. Após análise de bibliografia sobre o tema, como por exemplo do livro Coles (2001), do livro Beirlant, Goegebeur, Segers e Teugels (2006) e do artigo Penalva, Neves e Nunes (2013), foi verificado que os conceitos de que se necessitava para efetuar este estudo são comuns à bibliografia. Sendo este modelo teórico, instrumental para o estudo, visto que se trata da aplicação da Teoria dos Valores Extremos a uma situação concreta – o estudo do fluxo do tráfego diário na Ponte 25 de Abril – procurou-se identificar a bibliografia que mais facilmente e de modo acessível apresentava este modelo. Assim foi selecionada como fonte principal, para a apresentação deste modelo, o Livro *An Introduction to Statistical Modeling of Extreme Values*, de Coles (2001). Este capítulo é essencialmente constituído por um resumo da estrutura e da teoria apresentada no Capítulo 1 ao 4 do referido livro, podendo, por isso, não ser feita a referenciação bibliográfica convencional, típica de uma transcrição de textos referidos.

Na Análise de Valores Extremos, tal como o nome indica, faz-se a análise e a estimação da probabilidade de ocorrerem acontecimentos eventualmente mais extremos do que qualquer outro que já tenha sido anteriormente observado. O que realmente a distingue de qualquer outra análise é o facto de procurar quantificar o comportamento estocástico de um acontecimento que possui valores excecionalmente superiores ou inferiores aos valores mais usuais.

Quando não existem diretrizes empíricas ou físicas com as quais se formulam regras de extrapolação, os modelos utilizados são derivados de argumentos assintóticos. Supondo que se denota por $X_1, X_2, \ldots$ a sucessão do número de veículos que passam diariamente numa ponte, então

$$M_n = \max\{X_1, \ldots, X_n\}$$

representa o valor máximo diário de veículos durante um período de n observações. Poderia calcular-se de forma exata a distribuição de M_n, caso se conhecesse o comportamento estatístico de X_i. Como esse comportamento é desconhecido, esse cálculo não é possível. Contudo, o comportamento aproximado de M_n, para grandes valores de n, segue argumento de limites detalhados, permitindo $n \to \infty$, o que leva a uma família de modelos que podem ser ajustados pelos valores observados de M_n.

O paradigma de valor extremo pode ser a denominação da análise de Valores Extremos, visto que possui um princípio para a extrapolação de modelos baseada na implementação de limites matemáticos como aproximações de nível finito. É relevante que as limitações que estão implícitas na adoção do paradigma do valor extremo sejam tidas em conta: primeiro, é preciso ter cuidado ao tratar como resultados exatos os resultados obtidos através de argumentos assintóticos por detrás da elaboração dos modelos para amostras de dimensão finita; segundo, podem não ser razoáveis para um processo em estudo, as circunstâncias idealizadas que estão na base dos modelos que são derivados; depois, quando os modelos são implementados na prática, pode haver um desperdício de informações. Por exemplo, ao registar-se unicamente o máximo anual e a partir de vários máximos anuais chegar-se a um modelo que descreva as variações de um ano para outro, pode acontecer que em qualquer ano particular, existam eventos extremos adicionais que sejam mais extremos que outros valores de máximos anuais. Mas como não são o máximo desse ano acabam por ser excluídos da análise. Por isso, usam-se mais dados por ano no modelo estatísticos das r maiores observações e no Modelo Generalizado de Pareto usam-se as observações que se encontram acima deste.

A implementação estatística, como complemento ao desenvolvimento de modelos adequados para os valores extremos, é bastante relevante. E para que esta seja elaborada do melhor modo, ter-se-ão em consideração as seguintes observações: o método de estimação explorado será baseado nas técnicas da função de verosimilhança já que são únicas na capacidade que possuem de se adaptar à modificação do modelo, pois o método de estimação é o meio pelo qual os parâmetros desconhecidos de um modelo são diferidos com base em dados históricos; a quantificação da incerteza é importante dada a variabilidade da amostragem; os diagnósticos do modelo para avaliar a qualidade do ajuste do modelo têm a sua relevância; em relação ao uso de informações, são explorados modelos que usem vários dados, como os modelos multivariados, ou usam-se informações covariáveis ou se incorporam fontes adicionais de conhecimento numa análise.

3.2 Noções básicas de modelação estatística

3.2.1 Introdução

Nesta secção serão apresentadas algumas noções básicas de modelação estatística, se forem denotados por $x_1, \ldots, x_n$ os dados de uma sucessão do número de veículos observados diariamente. Logo, na travessia da ponte o tráfego no dia i tem a quantidade

aleatória de veículos, X_i. Quando o valor passa a ser conhecido é representado por x_i.

Supõe-se que X_i tem uma distribuição de probabilidade que atribui vários valores que o X_i possa ter. Os dados, $x_1,\ldots,x_n$ são um registo completo do padrão de tráfego que realmente existiu. Mas o papel da estatística não é apenas resumir o que já aconteceu, mas inferir as características da aleatoriedade no processo que gerou os dados.

As estatísticas consideram a sucessão $x_1,\ldots,x_n$ como realizações da sucessão de v.a.'s $X_1,\ldots,X_n$ e utilizam os dados para estimar a estrutura probabilística dessas v.a.'s.

3.2.2 Processos Aleatórios

Um **processo aleatório** é uma sucessão de v.a.'s $X_1,\ X_2,\ \ldots$. O exemplo mais simples é o de uma sucessão de v.a.'s i.i.d., que poderá ser, por exemplo, a descrição de fenómenos da vida real como cheias nos rios, picos de tráfego demasiado acentuado, e não só.

Definição 1. *Um processo aleatório $X_1,\ X_2,\ \ldots$ é considerado **estacionário** se, dado qualquer conjunto de inteiros $\{i_1,\ldots,i_k\}$ e qualquer número inteiro m, as distribuições conjuntas de $(X_{i_1},\ldots,X_{i_k})$ e de $(X_{i_{1+m}},\ldots,X_{i_{k+m}})$ forem idênticas.*

O que implica a estacionariedade é que, dado qualquer subconjunto de variáveis, a distribuição conjunta do mesmo subconjunto visto em m pontos de tempo permanece inalterada. Ao contrário de uma série independente, a estacionariedade não impede que X_i dependa de valores anteriores, embora X_{i+m} deva ter a mesma dependência dos seus valores anteriores.

3.2.3 Leis Limite

Definição 2. *Sendo $X_1,\ X_2,\ \ldots$ uma sucessão de v.a.'s, tendo respetivamente f.d. $F_1,\ F_2,\ \ldots$, diz-se que a sucessão **converge em distribuição** para a v.a. X, e escreve-se $X_n \xrightarrow{d} X$, tendo a f.d. F se*

$$F_n(x) \to F(x) \quad com \quad n \to \infty,$$

em todos os pontos de continuidade de F.

A utilidade de estabelecer uma distribuição limite F para uma sucessão de v.a.'s $X_1,\ X_2,\ \ldots$, para aplicações estatísticas, é justificar o uso de F como uma aproximação para a distribuição de X_n para n grande.

O **Teorema Limite Central (TLC)** é descrito de seguida.

Teorema 1. *Seja $X_1,\ X_2,\ \ldots$ uma sucessão de v.a.'s i.i.d. com média μ finita e variância σ^2 positiva. Então, definindo*

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n},$$

tem-se

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} Z \tag{3.1}$$

com $n\to\infty$, onde $Z\sim N(0,1)$.

Em aplicações estatísticas, o TLC é usado por interpretação de (3.1) como uma aproximação para a distribuição da média da amostra $\overline{X}_n$ para n grande. Isto é,

$$\overline{X}_n \dot{\sim} N(\mu, \sigma^2/n) \tag{3.2}$$

onde a notação $\dot{\sim}$ significa "é aproximadamente distribuído". O que faz o TLC notável é que a distribuição aproximada da média amostral é normal, independentemente da distribuição da sucessão de v.a.'s X_1, X_2, ...

3.2.4 Modelação Paramétrica

3.2.4.1 A Estrutura Paramétrica

A utilização de informações da amostra para fazer inferências sobre a estrutura da probabilidade da população, da qual os dados surgiram, é um objetivo comum na modelação estatística. No caso mais simples, os dados $x_1, \ldots, x_n$ são considerados realizações independentes da distribuição da população. A inferência equivale à estimativa dessa distribuição, para a qual existem duas abordagens: a paramétrica e não paramétrica. Na abordagem paramétrica é necessário, em primeiro lugar, adotar uma família de modelos dentro da qual a verdadeira distribuição dos dados esteja supostamente presente. Um modelo é escolhido, normalmente, por motivos empíricos, usando técnicas exploratórias para verificar famílias de modelos que parecem amplamente consistentes com os dados disponíveis. Outra hipótese, é utilizar as leis limite como aproximações. Já se mencionou no contexto de se usar a distribuição normal, como uma aproximação da distribuição das médias amostrais, e a abordagem também é central para o desenvolvimento de modelos de valores extremos.

Na discussão subsequente, restringiu-se a abordagem ao caso de uma variável aleatória (v.a.) contínua cuja função de densidade de probabilidade (f.d.p.) existe, apesar dos argumentos se aplicarem mais amplamente. Também se assumem que os dados $x_1, \ldots, x_n$ representam realizações independentes de uma v.a. X cuja a f.d.p. pertence a uma família de distribuições de probabilidade com funções de densidade $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ conhecida. Denota-se o verdadeiro valor do parâmetro θ por θ_0. A inferência é reduzida à estimativa do parâmetro θ_0 dentro do espaço de parâmetros Θ. O parâmetro θ pode ser um escalar, como $\theta = p$ na família binomial, ou pode representar um vetor de parâmetros, tal como $\theta = (\mu, \sigma)$ na família normal.

3.2.4.2 Estimação por Máxima Verosimilhança

Um método de estimação é a máxima verosimilhança (MV). Cada valor de $\theta \in \Theta$ define um modelo em $\mathcal{F}$ que atribui probabilidades diferentes aos dados observados, se as variáveis forem discretas. A probabilidade dos dados observados como uma função de θ é chamada função de verosimilhança. Os valores de θ que têm uma alta probabilidade de verosimilhança correspondem a modelos que dão uma probabilidade elevada

aos dados observados. O princípio da estimação por MV é adotar o modelo com maior verosimilhança, já que este é o que atribui maior probabilidade aos dados observados.

Em maior detalhe, referindo-se à situação em que $x_1, \ldots, x_n$ são realizações independentes de uma v.a. com f.d.p. $f(x;\theta_0)$, a **função de verosimilhança** é

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{3.3}$$

Lembrando que as variáveis $X_1, \ldots, X_k$ são **mutuamente independentes se**

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = \prod_{i=1}^{k} f_{X_i}(x_i). \tag{3.4}$$

A factorização em (3.3) é devida então a (3.4) para observações independentes. Nestes casos é mais conveniente aplicar logaritmos e trabalhar com a **função log-verosimilhança**

$$\ell(\theta) = \log\, L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta). \tag{3.5}$$

O **estimador de MV** $\hat{\theta}_0$ de θ_0 é definido como o valor de θ que maximiza a função de verosimilhança apropriada. Uma vez que a função logaritmo é monótona, a log-verosimilhança tem o seu máximo no mesmo ponto que a função de verosimilhança, pelo que o estimador de MV também maximiza a função log-verosimilhança correspondente.

3.2.4.3 Normalidade Aproximada do Estimador de Máxima Verosimilhança

Um benefício substancial da adoção da MV como princípio para a estimação de parâmetros é o facto de ser amplamente aplicável e estar disponível para várias distribuições de amostragem úteis. Isto leva a aproximações para erros padrão e Intervalos de confiança (IC). Destes obtêm-se alguns resultados úteis.

Cada um dos resultados é uma lei limite assintótica obtida à medida que o tamanho da amostra n tende para infinito. Os resultados são válidos apenas sob condições de regularidade, cuja precisão melhora à medida que n aumenta.

Teorema 2. *Sejam $x_1, \ldots, x_n$ realizações independentes de uma distribuição dentro de uma família paramétrica $\mathcal{F}$, sendo MVN a notação de uma distribuição Normal Multivariada, e $l(\bullet)$ e $\hat{\theta}_0$ denotam, respetivamente a função log-verosimilhança e o estimador de MV d-dimensional do modelo do parâmetro θ_0. Então, sob condições de regularidade para grandes n*

$$\hat{\theta}_0 \dot{\sim} MVN_d\left(\theta_0, I_E(\theta_0)^{-1}\right),$$

onde

$$I_E(\theta) = \begin{bmatrix} e_{1,1}(\theta) & \cdots & \cdots & e_{1,d}(\theta) \\ \vdots & \ddots & e_{i,j}(\theta) & \vdots \\ \vdots & e_{j,i}(\theta) & \ddots & \vdots \\ e_{d,1}(\theta) & \cdots & \cdots & e_{d,d}(\theta) \end{bmatrix},$$

com

$$e_{i,j}(\theta) = E\left\{-\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}(\theta)\right\}.$$

□

A matriz $I_E(\theta)$ é normalmente referida como a **matriz da informação esperada.**

O Teorema 2 pode ser usado para se obterem IC aproximados para componentes individuais de $\theta_0 = (\theta_1, \ldots, \theta_d)$. Denotando um termo arbitrário no inverso de $I_E(\theta)$ por $\psi_{i,j}$, decorre das propriedades da distribuição normal multivariada que, para n grande,

$$\hat{\theta}_i \dot{\sim} N(\theta_i, \psi_{i,i}).$$

Portanto, se $\psi_{i,i}$ fosse conhecido, um IC $(1-\alpha)\times 100\%$ aproximado para θ_i seria

$$\hat{\theta}_i \pm z_{\frac{\alpha}{2}} \sqrt{\psi_{i,i}} \tag{3.6}$$

onde $z_{\frac{\alpha}{2}}$ é o quantil $\left(1-\frac{\alpha}{2}\right)$ da distribuição normal padrão. Uma vez que o verdadeiro valor de θ_0 é habitualmente desconhecido, é comum aproximar os termos de I_E com os da **matriz de informação observada**, definida por

$$I_O(\theta) = \begin{bmatrix} -\frac{\partial^2 \ell}{\partial {\theta_1}^2}(\theta) & \cdots & \cdots & -\frac{\partial^2 \ell}{\partial\theta_1 \partial\theta_d}(\theta) \\ \vdots & \ddots & -\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}(\theta) & \vdots \\ \vdots & -\frac{\partial^2 \ell}{\partial\theta_j \partial\theta_i}(\theta) & \ddots & \vdots \\ -\frac{\partial^2 \ell}{\partial\theta_d \partial\theta_1}(\theta) & \cdots & \cdots & -\frac{\partial^2 \ell}{\partial {\theta_d}^2}(\theta) \end{bmatrix}$$

e avaliado em $\theta = \hat{\theta}$. Denotando os termos do inverso desta matriz por $\tilde{\psi}_{i,j}$, segue-se que um IC $(1-\alpha)$ aproximado para θ_i, é

$$\hat{\theta}_i \pm z_{\frac{\alpha}{2}} \sqrt{\tilde{\psi}_{i,i}}. \tag{3.7}$$

Estes intervalos são frequentemente mais precisos do que os obtidos em (3.6).

Embora uma família paramétrica $\mathcal{F}$ possa ser indexada por um parâmetro θ, no qual θ_0 representa o verdadeiro valor, pode não ser θ_0 o valor de particular interesse. Em vez disso, pode ser alguma função $\phi_0 = g(\theta_0)$ que se pretenda estimar, onde ϕ_0 pode ter uma dimensão diferente de θ_0. Restringe-se a atenção para a situação em que ϕ_0 é uma função escalar de θ_0. Isto é útil, muitas vezes, na modelação do valor extremo, onde θ_0 é o vetor do parâmetro de uma distribuição representante do comportamento do valor extremo, mas a probabilidade de algum acontecimento extremo – que é uma função de θ_0 – é que é necessária. Os dois resultados seguintes permitem que inferências de MV de θ_0 sejam transformadas para fornecer inferências correspondentes em ϕ_0.

Teorema 3. *Se $\hat{\theta}_0$ é a estimativa da MV de θ_0 e $\phi = g(\theta)$ é uma função escalar, então a estimativa de MV ϕ_0 é dada por $\hat{\phi}_0 = g(\hat{\theta}_0)$.* □

Este resultado significa que a estimativa de MV de qualquer função de θ_0 é obtida por substituição simples.

Teorema 4. *Seja $\hat{\theta}_0$ o estimador de MV da maior amostra do parâmetro d-dimensional θ_0 com matriz variância-covariância aproximada V_θ. Então se $\phi = g(\theta)$ é uma função escalar, o estimador de MV de $\phi_0 = g(\theta_0)$ satisfaz*

$$\hat{\phi}_0 \dot{\sim} N\left(\phi_0, V_\phi\right),$$

onde

$$V_\phi = \nabla\phi^T V_\theta \nabla\phi,$$

com

$$\nabla\phi = \left[\frac{\partial\phi}{\partial\theta_1}, \ldots, \frac{\partial\phi}{\partial\theta_d}\right]^T,$$

avaliado em $\hat{\theta}_0$. □

O Teorema 4 é conhecido como **método delta** e permite que a normalidade aproximada de $\hat{\theta}_0$ seja usada para obter IC para ϕ_0.

3.2.4.4 A Inferência Aproximada Utilizando a Função Desvio

O estimador de verosimilhança baseia-se na **função de desvio**, definida por

$$D(\theta) = 2\left\{\ell(\hat{\theta}_0) - \ell(\theta)\right\}. \tag{3.8}$$

Valores de θ com um desvio pequeno correspondem a modelos com alta verosimilhança. Deste modo, um critério natural para derivar regiões de confiança é especificar uma região de confiança

$$C = \{\theta : D(\theta) \leq c\}$$

para algumas escolhas de c. Como não é possível escolher c, de tal forma, que a região correspondente C tenha uma probabilidade pré-específica, $(1-\alpha)$, de conter o verdadeiro parâmetro θ_0, pois iria exigir que se conhecesse a distribuição da população, usa-se uma aproximação para a distribuição de amostragem que é válida para amostras de grandes dimensões.

Para o teorema seguinte é útil ter em conta a seguinte definição de distribuição para variáveis aleatórias contínuas.

Definição 3. *Se $Z_1, \ldots, Z_k$ são variáveis normais padronizadas independentes, a variável*

$$X = {Z_1}^2 + \cdots + {Z_k}^2$$

tem uma distribuição qui-quadrado com k graus de liberdade e escreve-se $X \sim \chi^2_k$.

Teorema 5. *Sejam $x_1, \ldots, x_n$, realizações independentes de uma distribuição dentro de uma família paramétrica $\mathcal{F}$, e $\hat{\theta}_0$ o estimador de MV do parâmetro θ_0 do modelo d-dimensional. Então, para n grande, sob condições de regularidade adequadas, a função de desvio (3.8) satisfaz*

$$D(\theta_0) \dot{\sim} \chi^2_d.$$

□

Segue do Teorema 5 que uma região de confiança $(1-\alpha)$ aproximada para θ_0 é dada por

$$C_\alpha = \{\theta : D(\theta) \leq c_\alpha\},$$

onde c_α é o quantil $(1-\alpha)$ da distribuição χ^2_d.

3.2.4.5 A Inferência Usando a Função de Verosimilhança de Perfil

Uma alternativa ao método que faz inferências numa componente particular θ_i de um vetor de parâmetros θ é o método baseado no perfil da verosimilhança. A log-verosimilhança para θ pode ser formalmente escrita como $\ell(\theta_i, \theta_{-i})$, onde θ_{-i}, denota todas as componentes de θ excluindo θ_i. O **perfil log-verosimilhança** para θ_i é definido como

$$\ell_p(\theta_i) = \max_{\theta_{-i}} \ell(\theta_i, \theta_{-i}).$$

Ou seja, para cada valor de θ_i, o perfil de log-verosimilhança é a log-verosimilhança maximizada em relação a todos os outros componentes de θ.

Esta definição pode ser generalizada para a situação onde θ pode ser dividido em duas componentes, $(\theta^{(1)}, \theta^{(2)})$, das quais $\theta^{(1)}$ é o vetor de dimensão k de interesse e $\theta^{(2)}$ corresponde aos componentes restantes $(d-k)$.

Teorema 6. *Sejam $x_1, \ldots, x_n$ realizações independentes de uma distribuição pertencente a uma família paramétrica $\mathcal{F}$, e $\hat{\theta}_0$ o estimador de verosimilhança máximo do parâmetro $\theta_0 = (\theta^{(1)}, \theta^{(2)})$ do modelo d-dimensional, onde $\theta^{(1)}$ é um subconjunto k-dimensional de θ_0. Então, sob condições de regularidade adequadas, para grandes valores de n*

$$D_p\left(\theta^{(1)}\right) = 2\left\{\ell\left(\hat{\theta}_0\right) - \ell_p(\theta^{(1)})\right\} \dot{\sim} \chi^2_k.$$

□

O Teorema 6 é frequentemente utilizado em duas situações diferentes. Primeiro, por um componente único θ_i, $C_\alpha = \left\{\theta_i : D_p(\theta_i) \leq c_\alpha\right\}$ é um IC $(1-\alpha)\times 100\%$ para θ_i, onde c_α é o quantil $(1-\alpha)$ da distribuição χ^2_1. A segunda aplicação é a seleção de modelos. Supondo que $\mathcal{M}_1$ é um modelo com o vetor de parâmetros θ, e o modelo $\mathcal{M}_0$ é o subconjunto do modelo $\mathcal{M}_1$, obtido restringindo k dos componentes de θ para ser, por exemplo, zero. Assim, θ pode ser partido em duas partes como $\theta = (\theta^{(1)}, \theta^{(2)})$, onde o primeiro componente, da dimensão k, é zero no modelo $\mathcal{M}_0$. Agora, se $\ell_1(\mathcal{M}_1)$ for a log-verosimilhança maximizada

para o modelo $\mathcal{M}_1$ e, sendo $\ell_0(\mathcal{M}_0)$ a log-verosimilhança maximizada para o modelo $\mathcal{M}_0$, define-se

$$D = 2\{\ell_1(\mathcal{M}_1) - \ell_0(\mathcal{M}_0)\}$$

como a **estatística de desvio**. Pelo Teorema 6, $C_\alpha = \left\{\theta^{(1)} : D_p\left(\theta^{(1)}\right) \leq c_\alpha\right\}$ compreende uma região de confiança $(1-\alpha)$ para o verdadeiro valor de $\theta^{(1)}$, onde D_p é o perfil de desvio e c_α é o quantil $(1-\alpha)$ da distribuição χ^2_k. Portanto, para verificar se $\mathcal{M}_0$ é uma redução plausível do modelo $\mathcal{M}_1$, é suficiente verificar se 0 está em C_α, que é equivalente a verificar se $D < c_\alpha$. Isto é denominado **teste de razão de verosimilhança**.

Teorema 7. *Seja $\mathcal{M}_0$ com o parâmetro $\theta^{(2)}$ o sub-modelo de $\mathcal{M}_1$ com o parâmetro $\theta_0 = (\theta^{(1)}, \theta^{(2)})$, sob a restrição de que o subvetor k-dimensional $\theta^{(1)} = 0$. Sejam $\ell_0(\mathcal{M}_0)$ e $\ell_1(\mathcal{M}_1)$ os valores maximizados da log-verosimilhança para os modelos $\mathcal{M}_0$ e $\mathcal{M}_1$, respetivamente. Um teste da validade do modelo $\mathcal{M}_0$ em relação ao $\mathcal{M}_1$, no nível de significância α, é rejeitar $\mathcal{M}_0$ em favor de $\mathcal{M}_1$, se $D = 2\{\ell_1(\mathcal{M}_1) - \ell_0(\mathcal{M}_0)\} > c_\alpha$, onde c_α é o quantil $(1-\alpha)$ da distribuição χ^2_k.* □

Finalmente, observa-se que é provável que cada uma das aproximações, de amostras de grandes dimensões é válida quando $x_1, ..., x_n$ são realizações independentes, mas não identicamente distribuídas de uma família indexada por um parâmetro θ.

3.2.4.6 Diagnóstico do Modelo

A razão pela qual se ajusta um modelo estatístico a dados é para tirar conclusões sobre algum aspeto da população da qual os dados foram extraídos. A questão principal diz respeito à capacidade do modelo para descrever variações na população em geral. A única opção que normalmente está disponível é julgar a precisão de um modelo em termos do seu acordo com os dados que foram realmente utilizados para estimar.

Assumindo que os dados $x_1, ..., x_n$ são realizações independentes de uma população com f.d. desconhecida F, uma estimativa de F, denotada por $\hat{F}$, é obtida pela MV, e quer-se avaliar a possibilidade da amostra ser proveniente de $\hat{F}$. Primeiro, uma estimativa do modelo de F pode ser obtida empiricamente a partir dos dados. Denotando por $x_{(1)}, \ldots, x_{(n)}$ a amostra ordenada, de modo que $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, sendo $\tilde{F}$ uma estimativa da verdadeira probabilidade de F e para qualquer um dos $x_{(i)}$, exatamente i das n observações têm um valor menor ou igual a $x_{(i)}$, então uma estimativa empírica da probabilidade de uma observação ser menor ou igual a $x_{(i)}$ é $\tilde{F}\left(x_{(i)}\right) = i/n$. Um ligeiro ajuste para $\tilde{F}\left(x_{(i)}\right) = i/(n+1)$ é geralmente feito para evitar ter $\tilde{F}\left(x_{(i)}\right) = 1$. Isto leva à seguinte definição.

Definição 4. *Dada uma amostra ordenada de observações independentes*

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

*de uma população com f.d. F, a **função de distribuição empírica** é definida por*

$$\tilde{F}(x) = \frac{i}{n+1} \text{ para } x_{(i)} \leq x < x_{(i+1)}.$$

Como $\tilde{F}$ é uma estimativa da verdadeira distribuição de probabilidade F, deverá estar de acordo com o modelo candidato, $\hat{F}$, desde que este seja uma estimativa adequada de F. Vários procedimentos da qualidade de ajuste são baseados nas comparações de $\tilde{F}$ e $\hat{F}$. Duas técnicas gráficas, em particular, são usadas frequentemente e descrevem-se de seguida.

Definição 5. *Dada uma amostra ordenada de observações independentes*

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

*de uma população com f.d. estimada $\hat{F}$, um **gráfico de probabilidade** consiste nos pontos*

$$\left\{\left(\hat{F}\left(x_{(i)}\right), \frac{i}{n+1}\right) : i = 1, \ldots, n\right\}.$$

*E um **gráfico de quantis** consiste nos pontos*

$$\left\{\left(\hat{F}^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right) : i = 1, \ldots, n\right\}.$$

Se $\hat{F}$ é um modelo razoável para a distribuição da população, os pontos do gráfico de probabilidade devem estar próximos da diagonal da unidade. Desvios substanciais da linearidade fornecem evidência de uma falha em $\hat{F}$ como um modelo para os dados.

Se $\hat{F}$ é uma estimativa razoável de F, então o gráfico quantil também deve consistir em pontos próximos à diagonal da unidade, ou seja, a $y = x$.

O gráfico de probabilidade e o gráfico de quantis contêm as mesmas informações expressas numa escala diferente. No entanto, a percepção que é ganha em diferentes escalas pode ser importante.

3.3 Teoria Clássica e modelos dos Valores Extremos

3.3.1 Modelos Assintóticos

3.3.1.1 Formulação do Modelo

O modelo que será apresentado é a pedra angular da teoria dos valores extremos. Este foca-se no comportamento estatístico de

$$M_n = max\{X_1, \ldots, X_n\}$$

onde $X_1, \ldots, X_n$ é uma sucessão de v.a.'s independentes com uma f.d. comum, F. Em aplicações, o X_i geralmente representa valores de um processo medido numa escala de tempo regular, de modo que M_n representa o máximo do processo em n unidades de observação. Se n é o número de observações num ano, então M_n corresponde ao máximo anual.

Em teoria, a distribuição de M_n pode ser derivada exatamente para todos os valores de n:

$$\Pr\{M_n \leq z\} = \Pr\{X_1 \leq z, \ldots, X_n \leq z\} = \Pr\{X_1 \leq z\} \times \cdots \times \Pr\{X_n \leq z\} = \{F(z)\}^n . \qquad (3.9)$$

No entanto, a f.d. F é desconhecida, logo isto não é imediatamente útil na prática. Uma possibilidade é utilizar técnicas estatísticas padrão para estimar F a partir de dados observados, e depois substitui-se a estimativa em (3.9). Infelizmente, discrepâncias muito pequenas na estimativa de F podem levar a discrepâncias substanciais para F^n.

Uma abordagem alternativa é aceitar que F é desconhecida e procurar famílias de modelos aproximadas de F^n, que podem ser estimados com base apenas nos dados extremos. Isto é semelhante à prática habitual de aproximar a distribuição das médias amostrais pela distribuição normal, como justificado pelo TLC.

Observa-se o comportamento de F^n com $n \to \infty$. Mas isso simplesmente não é suficiente: para qualquer $z < z_+$, onde z_+ é o limite superior do suporte de F, $F^n(z) \to 0$ com $n \to \infty$, pelo que a distribuição de M_n será degenerada com massa de probabilidade concentrada em z_+. Esta dificuldade é evitada permitindo uma normalização linear da variável M_n:

$$M_n{}^* = \frac{M_n - b_n}{a_n},$$

para sucessões de constantes $a_n > 0$ e b_n. Escolhas apropriadas de a_n e b_n estabilizam a localização e a escala de $M_n{}^*$ à medida que n aumenta, evitando as dificuldades que surgem com a variável M_n. Por isso, procuram-se distribuições de limites para $M_n{}^*$, com escolhas apropriadas de a_n e b_n, em vez de M_n.

3.3.1.2 Teorema dos Modelos Extremos

Toda a gama de distribuições de limites possíveis para $M_n{}^*$ é dada pelo Teorema 8, **o Teorema dos Modelos Extremos**.

Teorema 8. *Se existirem sucessões reais $\{a_n > 0\}$ e $\{b_n\}$ de tal modo que*

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z) \quad com \quad n \to \infty, \tag{3.10}$$

onde G é uma f.d. não-degenerada, então G é uma das seguintes distribuições:

I:

$$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \ -\infty < z < \infty\ ; \tag{3.11}$$

II:

$$G(z) = \begin{cases} 0, & z \leq b, \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b; \end{cases} \tag{3.12}$$

III:

$$G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{-\alpha}\right]\right\}, & z < b, \\ 1, & z \geq b; \end{cases} \tag{3.13}$$

*para os parâmetros $a > 0$, $b \in \mathbb{R}$ e, no caso das distribuições **II** e **III**, $\alpha > 0$.* □

Portanto, o Teorema 8 afirma que os máximos da amostra $\frac{M_n - b_n}{a_n}$ reescalonados convergem na distribuição para uma variável com uma distribuição que se encontra dentro de uma das famílias apresentadas. Estas três classes de distribuições são denominadas **Distribuições de Valores Extremos**, com tipos **I**, **II** e **III**, amplamente conhecidas como as famílias de **Gumbel**, de **Fréchet** e de **Weibull**, respetivamente. Cada família tem um parâmetro de localização, b, e de escala, a; para além disso, as famílias **Fréchet** e **Weibull** têm um parâmetro de forma, α.

O Teorema 8 implica que, quando M_n pode ser estabilizado com sucessões adequadas a_n e b_n, a variável normalizada correspondente $M_n{}^*$ tem uma distribuição limite que deve ser um dos três tipos de distribuições de valores extremos. A característica notável deste resultado é que os três tipos de distribuições de valor extremo são os únicos limites possíveis para a distribuição $M_n{}^*$, independentemente da distribuição F para a população. É neste sentido, que o teorema fornece uma distribuição limite análoga ao TLC.

3.3.1.3 Distribuição generalizada dos valores Extremos

Os três tipos de limites que surgem no Teorema 8 têm formas distintas de comportamento, correspondendo às diferentes formas do comportamento da cauda da f.d. F do X_i. Para tornar esta ideia mais clara, considere-se o comportamento da distribuição limite G em z_+, no seu limite superior do suporte. Para a distribuição de **Weibull** z_+ é finita, enquanto que para as restantes distribuições $z_+ = \infty$. No entanto, a densidade de G decai exponencialmente para a distribuição de **Gumbel** e polinomialmente para a distribuição de **Fréchet**, correspondendo a taxas relativamente diferentes de quedas na cauda de F. Segue que nas aplicações, as três diferentes distribuições dão uma representação um pouco distinta do comportamento do valor extremo. Nas primeiras aplicações da teoria de valores extremos era comum adotar uma das três famílias e depois estimar os parâmetros relevantes dessa distribuição. Mas existem dois pontos fracos: primeiro, é necessária uma técnica para escolher qual das três famílias é mais apropriada para os dados em questão; segundo, uma vez tomada tal decisão, as inferências subsequentes assumem que esta escolha é a correta e não têm em consideração a incerteza que tal seleção envolve, embora essa incerteza possa ser substancial.

Reformulando os modelos do Teorema 8 é possível uma melhor análise. É fácil verificar que as famílias **Gumbel**, **Fréchet** e **Weibull** podem ser combinadas numa única família de modelos tendo a função de distribuição da forma

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{3.14}$$

definido no conjunto $\{z : 1 + \xi\left(\frac{z-\mu}{\sigma}\right) > 0\}$, onde os parâmetros satisfazem $-\infty < \mu < \infty$, $\sigma > 0$ e $-\infty < \xi < \infty$. Esta é a família do **valor extremo generalizado (GEV – *generalized extreme value*)** das distribuições. O modelo tem três parâmetros: um parâmetro de localização, μ; um parâmetro de escala, σ; e um parâmetro de forma, ξ. As classes do tipo **II** e do tipo **III** da distribuição de valores extremos correspondem, respetivamente, aos

casos $\xi > 0$ e $\xi < 0$ nesta parametrização. O subconjunto da família GEV com $\xi = 0$ é interpretado como o limite de (3.14) quando $\xi \to 0$, levando à **família Gumbel** com f.d.

$$G(z) = \exp\left[-\exp\left\{-\left(\frac{z-\mu}{\sigma}\right)\right\}\right], \quad -\infty < z < \infty.$$

A unificação das três famílias numa única família simplifica muito a implementação estatística. Através da inferência em ξ, os próprios dados determinam o tipo mais adequado de comportamento da cauda, e não há necessidade de fazer julgamentos subjetivos a priori sobre qual a distribuição de valor extremo individual a adotar. Além disso, a incerteza no valor inferido de ξ, mede a falta de certeza sobre qual dos três tipos de modelos originais é o mais apropriado para um determinado conjunto de dados.

Pode-se assim reformular o Teorema 8.

Corolário 1. *Sendo G um membro da família GEV então*

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\},$$

definido no conjunto $\left\{z\colon 1+\xi\left(\frac{z-\mu}{\sigma}\right) > 0\right\}$, *onde* $-\infty < \mu < \infty$, $\sigma > 0$ *e* $-\infty < \xi < \infty$. □

Interpretando o limite no Corolário 1, como uma aproximação para grandes valores de n, é recomendado o uso da família GEV para modelar a distribuição de máximos de grandes sucessões. A aparente dificuldade pelo facto das constantes de normalização serem desconhecidas, na prática, é facilmente resolvido. Assumindo (3.14),

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \approx G(z)$$

para um n suficientemente grande. De modo equivalente, tem-se

$$\Pr\{M_n \leq z\} \approx G\left\{\frac{z-b_n}{a_n}\right\} = G^*(z),$$

onde G^* é outro membro da família GEV. Por outras palavras, se o Corolário 1 permite a aproximação da distribuição de $M_n{}^*$ por um membro da família GEV para n grandes, a distribuição do próprio M_n, também pode ser aproximada, por um membro diferente da mesma família. Uma vez que, os parâmetros da distribuição têm que ser estimados de qualquer forma, é irrelevante, na prática, que os parâmetros da distribuição G sejam diferentes daqueles de G^*.

Este argumento leva à seguinte abordagem para modelar os extremos de uma série de observações independentes $X_1, X_2, \ldots$. Os dados são agrupados em blocos em sucessões de observações de comprimento n, para algum valor grande de n, gerando uma série de blocos de máximos, $M_{n,1}, \ldots, M_{n,m}$, para os quais a distribuição GEV pode ser ajustada. Frequentemente, os blocos são escolhidos para corresponder a um período de tempo de comprimento de um ano, que nestes casos significa que n é o número de observações num

ano e os máximos dos blocos são máximos anuais. Estimativas de quantis extremos, da distribuição máxima anual, são obtidos invertendo a equação (3.14)

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], & \text{para } \xi \neq 0, \\ \mu - \sigma \log\{-\log(1-p)\}, & \text{para } \xi = 0, \end{cases} \tag{3.15}$$

onde $G(z_p) = 1 - p$. Na terminologia comum, z_p é o **nível de retorno (NR)** associado ao **período de retorno** $\frac{1}{p}$, porque com um grau razoável de precisão, o nível z_p é esperado que seja excedido, em média, uma vez a cada $\frac{1}{p}$ anos. Mais precisamente, z_p é excedido pelo máximo anual, em qualquer ano, com probabilidade p.

Como os quantis permitem que os modelos de probabilidade sejam expressos numa escala de dados, a relação do modelo GEV com os seus parâmetros é mais fácil de interpretar em termos de expressões de quantis (3.15). Em particular, definindo $y_p = -\log(1-p)$, de modo que

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - y_p^{-\xi}\right], & \text{para } \xi \neq 0, \\ \mu - \sigma \log y_p\,, & \text{para } \xi = 0; \end{cases}$$

segue-se que, se z_p é traçado contra y_p, numa escala logarítmica – ou equivalente, se z_p é traçado contra $\log y_p$ – o gráfico é linear no caso de $\xi = 0$. Se $\xi < 0$ o gráfico é convexo com limite assintótico com $p \to 0$ em $\mu - \frac{\sigma}{\xi}$; se $\xi > 0$ o gráfico é côncavo e não tem limite finito. Este gráfico na figura 3.1 é uma representação do **gráfico do nível de retorno**. Devido à simplicidade de interpretação, e tendo presente que a escolha de escala comprime a cauda da distribuição, de modo que o efeito da extrapolação é realçado, os gráficos de NR são particularmente convenientes para a apresentação e a validação do modelo. A figura 3.1 mostra gráficos de NR para uma gama de parâmetros de forma, retirada do livro Coles (2001).

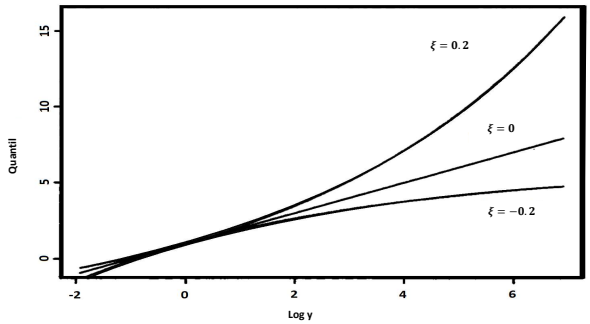


Figura 3.1: Gráficos de NR da distribuição GEV com parâmetros de forma $\xi = -0.2$, $\xi = 0$ e $\xi = 0.2$, respetivamente

3.3.2 Inferência para a distribuição GEV

3.3.2.1 Considerações Gerais

Motivado pelo Corolário 1, o GEV fornece um modelo para a distribuição de blocos de máximos. A aplicação consiste em agrupar os dados em blocos de igual comprimento, e ajustar o GEV ao conjunto de blocos de máximos. Mas ao implementar este modelo para qualquer conjunto de dados, em particular, a escolha do tamanho do bloco pode ser crítica. A escolha equivale a uma troca entre viés e variância: os blocos que são muito pequenos significam que a aproximação pelo modelo limite no Corolário 1 é provavelmente pobre, levando a um enviesamento na estimativa e extrapolação; blocos grandes geram poucos blocos de máximos, levando a uma grande variância na estimação. Por isso, considerações pragmáticas muitas vezes levam à adoção de blocos de duração de um ano. Por exemplo, se apenas os dados máximos anuais tiverem sido gravados, então, o uso de blocos mais curtos não é uma opção. Até quando este não é o caso, é provável que uma análise dos dados máximos anuais seja mais robusta do que uma análise baseada em blocos mais curtos levando a que as condições do Corolário 1 não sejam respeitados. Por exemplo, as temperaturas diárias, é provável, que variem consoante a estação, opondo-se à suposição de que X_i tenha uma distribuição comum. Se os dados foram agrupados em blocos de aproximadamente três meses, o máximo do bloco do verão provavelmente será muito maior do que o bloco de inverno, e uma inferência que não conseguiu levar esta não-homogeneidade em conta poderia dar resultados imprecisos. Fazendo, em vez disso, blocos de comprimento de um ano significa que a suposição de que o bloco de máximos tem uma distribuição comum é plausível, embora a justificação formal para a aproximação do GEV permanece inválida.

Agora simplifica-se a notação denotando os blocos de máximos por $Z_1, \ldots, Z_m$. Estes são assumidos como variáveis independentes de uma distribuição GEV cujos parâmetros devem ser estimados. Se os X_i forem independentes, então os Z_i, também serão independentes. No entanto, a independência do Z_i é provável que seja uma aproximação razoável, mesmo se X_i constituir uma série dependente. Neste caso, embora não seja abrangido pelo Corolário 1, a conclusão de que o Z_i tem uma distribuição GEV ainda pode ser razoável.

Muitas técnicas têm sido propostas para a estimação de parâmetros em modelos de valor extremo. Cada técnica tem os seus prós e contras, mas a utilidade geral e a adaptabilidade à construção complexa de modelos de técnicas de verosimilhança básicas, tornam esta abordagem particularmente atraente.

Uma dificuldade potencial com o uso de métodos de verosimilhança para o GEV refere-se à validade das condições de regularidade, exigidas pelas propriedades assintóticas usuais, associadas ao estimador de MV. Tais condições não são satisfeitas pelo modelo GEV, porque os pontos finais da distribuição GEV são funções dos valores dos parâmetros: $\mu - \sigma/\xi$ é um limite superior do suporte da distribuição quando $\xi < 0$ e um ponto final inferior quando $\xi > 0$. Esta transgressão das condições usuais de regularidade, significa que os resultados da verosimilhança assintótica padrão, não são automaticamente aplicáveis.

Smith (1985) estudou este problema em detalhe e obteve os seguintes resultados:

- quando $\xi > -0.5$, os estimadores de MV são regulares, ou seja, têm as propriedades assintóticas usuais;
- quando $-1 < \xi < -0.5$, os estimadores de MV são geralmente obtidos, mas não possuem as propriedades assintóticas padrão;
- quando $\xi < -1$, os estimadores de probabilidade MV são improváveis de serem obtidos.

O caso $\xi \leq -0.5$ corresponde a distribuições com um limite muito curto da cauda superior. Esta situação raramente é encontrada em aplicações de modelação de valor extremo, por isso as limitações teóricas da abordagem da MV, geralmente, na prática não são um obstáculo.

3.3.2.2 Estimação por máxima Verosimilhança

Sob a suposição de que $Z_1, \ldots, Z_m$ sejam variáveis independentes com distribuição GEV, a log-verosimilhança para os parâmetros GEV quando $\xi \neq 0$ é

$$\ell(\mu, \sigma, \xi) = -m \log \sigma \; - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m} \left[1 + \xi \left(\frac{z_i - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}, \tag{3.16}$$

providencia-se que

$$1 + \xi \left(\frac{z_i - \mu}{\sigma}\right) > 0, \text{ para } i = 1, \ldots, m. \tag{3.17}$$

Em combinações de parâmetros para os quais (3.17) não é respeitado, correspondendo a uma configuração para a qual, pelo menos, um dos dados observados está além de um ponto final da distribuição, a verosimilhança é zero e a log-verosimilhança é igual a $-\infty$.

O caso $\xi = 0$ requer um tratamento separado usando o limite de Gumbel da distribuição GEV. Isto leva à log-verosimilhança

$$\ell(\mu, \sigma) = -m \log \sigma \; - \sum_{i=1}^{m} \left(\frac{z_i - \mu}{\sigma}\right) - \sum_{i=1}^{m} \exp \left\{-\left(\frac{z_i - \mu}{\sigma}\right)\right\}. \tag{3.18}$$

A Maximização do par de Equações (3.16) e (3.18) em relação ao vetor de parâmetros (μ, σ, ξ), leva ao estimador de MV com respeito a toda a família GEV. Não há solução analítica, mas para qualquer conjunto de dados a maximização é simples usando algoritmos de otimização numérica padrão. É necessário ter algum cuidado para garantir que tais algoritmos não se alterem, de tal modo, que fiquem combinações de parâmetros que não respeitem a (3.17), e também devem ser evitadas dificuldades numéricas que possam surgir da avaliação de (3.16) nas vizinhanças de $\xi = 0$. Este último problema resolve-se facilmente utilizando a (3.18) no lugar de (3.16) para os valores de ξ ficarem dentro de uma pequena janela à volta de zero.

Estando ξ sujeito às limitações discutidas anteriormente, a distribuição aproximada de $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ é normal multivariada com média (μ, σ, ξ) e a matriz de variância-covariância igual ao inverso da matriz de informação observada, avaliada na estimativa da MV. Embora esta matriz possa ser calculada analiticamente, é mais fácil usar técnicas de diferenciação numérica para avaliar as segundas derivadas e rotinas standard numéricas para realizar a inversão. IC e outras formas de inferência seguem imediatamente da normalidade aproximada do estimador.

3.3.2.3 Inferências para níveis de retorno

Por substituição das estimativas de MV dos parâmetros GEV em (3.15), a estimativa da MV de z_p para $0 < p < 1$, o NR $1/p$, é obtida da seguinte forma

$$\hat{z}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left[1 - y_p{}^{-\hat{\xi}}\right], \text{ para } \hat{\xi} \neq 0, \\ \hat{\mu} - \hat{\sigma}\log y_p \text{ , para } \hat{\xi} = 0, \end{cases} \tag{3.19}$$

onde $y_p = -\log(1-p)$. Além disso, pelo método delta,

$$Var\left(\hat{z}_p\right) \approx \nabla z_p{}^T V \nabla z_p, \tag{3.20}$$

onde V é a matriz de variância-covariância de $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ e

$$\nabla z_p{}^T = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right] = \left[1,\ -\xi^{-1}\left(1 - y_p{}^{-\xi}\right), \sigma\xi^{-2}\left(1 - y_p{}^{-\xi}\right) - \sigma\xi^{-1} y_p{}^{-\xi}\log y_p\ \right]$$

avaliado em $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$.

Geralmente são longos períodos de retorno, correspondendo a pequenos valores de p, que são de maior interesse. Se $\hat{\xi} < 0$ também é possível fazer inferências sobre o limite superior do suporte da distribuição, que é efetivamente o 'período infinito de retorno da observação', correspondendo, a z_p com $p = 0$. A estimativa da MV é

$$\hat{z}_0 = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}},$$

e (3.20) ainda é válida com

$$\nabla z_0{}^T = \left[1, \xi^{-1}, \sigma\xi^{-2}\right],$$

novamente avaliado em $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$. Quando $\hat{\xi} \geq 0$ a estimativa da MV do limite superior do suporte é infinita.

É necessário ter cuidado na interpretação das inferências do NR, especialmente para NR correspondentes a longos períodos. Primeiro, a aproximação normal da distribuição do estimador da MV pode ser pobre. Melhores aproximações são geralmente obtidas a partir do perfil adequado da função de verosimilhança. Fundamentalmente, as estimativas e as suas medidas de precisão baseiam-se no pressuposto de que o modelo está correto. Embora o modelo GEV seja apoiado por argumentos matemáticos, o seu uso na extrapolação é baseado em premissas não verificáveis, e as medidas de incerteza sobre os NR devem ser apropriadamente consideradas como limites inferiores que poderiam ser muito maiores se a incerteza devido à correção do modelo fosse tida em consideração.

3.3.2.4 O Perfil da Verosimilhança

A avaliação numérica, do perfil da verosimilhança para qualquer um dos parâmetros, individualmente, μ, σ ou ξ, é simples. Por exemplo, para obter o perfil da verosimilhança para ξ, fixa-se $\xi = \xi_0$, e maximiza-se a log-verosimilhança (3.16) em relação aos parâmetros restantes, μ e σ. Isto é repetido para um intervalo de valores de ξ_0. Os valores maximizados correspondentes da log-verosimilhança constituem o perfil log-verosimilhança para ξ, a partir do qual o Teorema 6 permite obter IC aproximados.

Esta metodologia também pode ser aplicada quando a inferência é necessária em algumas combinações de parâmetros. Em particular, podem-se obter IC para qualquer NR específico z_p. Isto requer uma reparametrização do modelo GEV, de modo que z_p seja um dos parâmetros do modelo, após o qual o perfil log-verosimilhança é obtido pela maximização em relação aos parâmetros restantes na maneira usual. A reparametrização é direta:

$$\mu = z_p + \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], \tag{3.21}$$

de tal modo, que a substituição de μ em (3.16) por (3.21) tenha o efeito desejado de expressar o modelo GEV em termos dos parâmetros (z_p, σ, ξ).

3.3.2.5 Verificação do Modelo

Embora seja impossível verificar a validade de uma extrapolação baseada num modelo GEV, a avaliação pode ser feita com referência aos dados observados. Isto não é suficiente para justificar a extrapolação, mas é um pré-requisito razoável.

Como descrito anteriormente, um gráfico de probabilidade é uma comparação entre funções de distribuição empírica e ajustada. Com os blocos de máximos ordenados deste modo $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(m)}$, a f.d. empírica avaliada em $z_{(i)}$ é dada por

$$\tilde{G}\left(z_{(i)}\right) = \frac{i}{m+1}.$$

Por substituição de estimativas de parâmetros em (3.14), as estimativas baseadas em modelos correspondentes são

$$\hat{G}\left(z_{(i)}\right) = \exp\left\{-\left[1 + \hat{\xi}\left(\frac{z_{(i)} - \hat{\mu}}{\hat{\sigma}}\right)\right]^{-\frac{1}{\hat{\xi}}}\right\}.$$

Se o modelo GEV estiver a funcionar bem,

$$\hat{G}\left(z_{(i)}\right) \approx \tilde{G}\left(z_{(i)}\right)$$

para cada i, então um gráfico de probabilidade constituído pelos pontos

$$\left\{\left(\tilde{G}\left(z_{(i)}\right), \hat{G}\left(z_{(i)}\right)\right),\ i = 1, \ldots, m\right\},$$

deve ficar perto da diagonal unidade. Quaisquer desvios substanciais da linearidade são indicativos de alguma falha no modelo GEV.

Uma fraqueza do gráfico de probabilidade para modelos de valor extremo é que ambos $\hat{G}\left(z_{(i)}\right)$ e $\tilde{G}\left(z_{(i)}\right)$ são obrigados a aproximar-se de 1 quando $z_{(i)}$ aumenta, enquanto é geralmente a precisão do modelo para grandes valores de z que é de maior preocupação. Ou seja, o gráfico de probabilidade fornece a menor informação na região de maior interesse. Esta falha é evitada pelo gráfico quantil, consistindo nos pontos

$$\left\{\left(\hat{G}^{-1}\left(\frac{i}{m+1}\right),\ z_{(i)}\right),\ i=1,\ldots,m\right\}, \tag{3.22}$$

onde, de (3.19)

$$\hat{G}^{-1}\left(\frac{i}{m+1}\right)=\hat{\mu}-\frac{\hat{\sigma}}{\hat{\xi}}\left[1-\left\{-\log\left(\frac{i}{m+1}\right)\right\}^{-\hat{\xi}}\right].$$

Saídas da linearidade no gráfico de quantis também indicam falha do modelo.

Conforme discutido anteriormente, o gráfico do NR, que compreende um gráfico de

$$z_p=\mu+\frac{\sigma}{\xi}\left[1-\{-\log(1-p)\}^{-\xi}\right]$$

contra $y_p=-\log(1-p)$ numa escala logarítmica, é particularmente conveniente para interpretar modelos de valor extremo. A cauda da distribuição é comprimida, de modo que, as estimativas do NR para longos períodos de retorno sejam exibidas, enquanto a linearidade do gráfico no caso $\xi=0$ fornece uma linha de base, contra a qual se julga o efeito do parâmetro que fora estimado.

Como resumo de um modelo ajustado, o gráfico de NR consiste no *locus* dos pontos

$$\left\{\left(\log y_p\ ,\hat{z}_p\right):\ 0<p<1\right\},$$

onde $\hat{z}_p$ é a estimativa da MV de z_p. Os IC podem ser adicionados ao gráfico para aumentar a sua informação. Estimativas empíricas da função do NR, obtidas a partir dos pontos (3.22), também podem ser adicionadas, permitindo que o gráfico do NR seja usado como um diagnóstico de modelo. Se o modelo GEV é adequado aos dados, a curva baseada no modelo e as estimativas empíricas devem estar razoavelmente de acordo. Qualquer discordância substancial ou sistemática, após o adiantamento para o erro de amostragem, sugere uma inadequação do modelo GEV.

Os gráficos de probabilidade, de quantis e de NR são baseados numa comparação entre modelos base e estimativas empíricas da f.d.. Para completar, um diagnóstico equivalente é baseado na função de densidade, ou seja, é uma comparação da f.d.p., de um modelo ajustado, com um histograma dos dados.

3.3.3 Generalização do modelo: o modelo estatístico das r maiores observações

3.3.3.1 Formulação do Modelo

Uma dificuldade implícita em qualquer análise de valores extremos é a quantidade limitada de dados para a estimativa do modelo. Os extremos são escassos, por definição,

por isso, as estimativas dos modelos, especialmente de NR extremos, têm uma grande variação. Esta questão motivou a procura por caracterizações do comportamento do valor extremo, que permita a modelação de dados, que não sejam apenas através de blocos de máximos.

Existem duas caracterizações gerais bem conhecidas. Uma é baseada em excedências de um limite elevado, a outra baseia-se no comportamento das estatísticas das r maiores observações dentro de um bloco, para valores pequenos de r. Este estudo concentra-se num modelo estatístico das r maiores observações.

Supondo que $X_1, X_2, \ldots$ é uma sucessão de v.a.'s i.d.d., e objetivam caracterizar o comportamento do extremo X_i. Na Secção 3.3.1.3, obteve-se que a distribuição limite, com $n \to \infty$, de M_n, adequadamente redimensionada, é GEV. Primeiro estende-se este resultado para outras estatísticas de ordem extrema, definindo

$$M_n{}^{(k)} = k \text{ maior estatística ordinal de } \{X_1, \ldots, X_n\},$$

e identificando o comportamento limitante dessa variável, para k fixo, com $n \to \infty$. O seguinte resultado generaliza o Teorema 8.

Teorema 9. *Se houver sucessões de constantes, $\{a_n > 0\}$ e $\{b_n\}$, de tal modo que*

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z) \quad \text{com } n \to \infty.$$

Para alguma f.d. não-degenerada G, tal que G é a f.d. GEV dada por (3.14), então, para um k fixo,

$$\Pr\left\{\frac{M_n{}^{(k)} - b_n}{a_n} \leq z\right\} \to G_k(z),$$

em que $\left\{z\colon 1 + \frac{\xi(z-\mu)}{\sigma} > 0\right\}$, onde

$$G_k(z) = \exp\{-\tau(z)\} \sum_{s=0}^{k-1} \frac{\tau(z)^s}{s!}, \tag{3.23}$$

com

$$\tau(z) = \left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}.$$

□

O Teorema 9 implica que, se a estatística das k maiores observações num bloco for normalizada exatamente da mesma maneira que o máximo, então a sua distribuição limite é da forma dada por (3.23), cujos parâmetros correspondem aos parâmetros da distribuição limite GEV do bloco máximo. Novamente, absorvendo as constantes de escala desconhecidas nos parâmetros de localização e de escala do modelo, segue-se que, para n grande, a distribuição aproximada de $M_n{}^{(k)}$ está dentro da família (3.23).

Há, no entanto, uma dificuldade ao usar (3.23) como modelo. A situação que ocorre muitas vezes, é de ter cada uma das r maiores observações dentro de cada um dos vários blocos, para alguns valores de r. Isto é, geralmente tem-se o vetor completo

$$M_n{}^{(r)} = \left(M_n{}^{(1)}, \ldots, M_n{}^{(r)}\right)$$

para cada um dos vários blocos. Enquanto o Teorema 9 dá uma família para a distribuição aproximada de cada um dos componentes de $M_n{}^{(r)}$, não dá a distribuição conjunta de $M_n{}^{(r)}$. Além disso, os componentes não podem ser independentes: $M_n{}^{(2)}$ não pode ser maior que $M_n{}^{(1)}$, por exemplo, logo o resultado de cada componente influencia a distribuição do outro. Consequentemente, o resultado do Teorema 9 não conduz em si mesmo a um modelo para $M_n{}^{(r)}$. Em vez disso, exige-se uma caracterização da distribuição conjunta limite de todo o vetor $M_n{}^{(r)}$. Com redimensionamento apropriado isto pode ser alcançado, mas a distribuição conjunta limite leva à intratabilidade. No entanto, o seguinte teorema dá a função densidade conjunta da distribuição limite.

Teorema 10. *Se houver sucessões de constantes, $\{a_n > 0\}$ e $\{b_n\}$, de tal modo, que*

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z)$$

com $n \to \infty$, para alguma f.d. não-degenerada G, então, para r fixo, a distribuição limite, com $n \to \infty$, de

$$\tilde{M}_n^{(r)} = \left(\frac{M_n{}^{(1)} - b_n}{a_n}, \ldots, \frac{M_n{}^{(r)} - b_n}{a_n}\right)$$

fica dentro da família com f.d.p. conjunta

$$f\left(z^{(1)}, \ldots, z^{(r)}\right) = \exp\left\{-\left[1 + \xi\left(\frac{z^{(r)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \times \prod_{k=1}^{r} \sigma^{-1}\left[1 + \xi\left(\frac{z^{(k)} - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}, \tag{3.24}$$

onde $-\infty < \mu < \infty$, $\sigma > 0$ e $-\infty < \xi < \infty$; $z^{(r)} \leq z^{(r-1)} \leq \cdots \leq z^{(1)}$; e $z^{(k)}$:$\xi\left(\frac{z^{(k)}-\mu}{\sigma}\right) > 0$ para $k = 1, \ldots, r$. □

No caso de $r = 1$, (3.24) reduz-se para a família de funções de densidade GEV. O caso $\xi = 0$ em (3.24) é interpretada como a forma limite com $\xi \to 0$, levando à família de funções de densidade

$$f\left(z^{(1)}, \ldots, z^{(r)}\right) = \exp\left\{-\exp\left[-\left(\frac{z^{(r)} - \mu}{\sigma}\right)\right]\right\} \times \prod_{k=1}^{r} \sigma^{-1}\exp\left[-\left(\frac{z^{(k)} - \mu}{\sigma}\right)\right], \tag{3.25}$$

para a qual o caso $r = 1$ reduz à densidade da família **Gumbel**.

3.3.3.2 Modelação das Estatísticas das r maiores observações

Tendo uma série de variáveis i.i.d., os dados são agrupados em m blocos. No bloco i as maiores observações r_i são gravadas, levando à série $M_i{}^{(r_i)} = \left(z_i{}^{(1)}, \ldots, z_i{}^{(r_i)}\right)$ para

$i = 1, \ldots, m$. É usual definir $r_1 = \cdots = r_m = r$ para algum valor de r específico, a não ser que menos dados estejam disponíveis em alguns blocos.

Assim como no modelo GEV a questão do tamanho do bloco equivale a uma troca entre viés e variância, o número "de ordem" das estatísticas usadas em cada bloco também: valores pequenos de r geram poucos dados o que leva a uma variância elevada; grandes valores de r são suscetíveis de não respeitar o suporte assintótico para o modelo, levando ao enviesamento. Na prática é comum selecionar o r_i maior possível, sujeito a diagnósticos de um modelo adequado.

A verosimilhança para este modelo é obtida a partir de (3.24) e (3.25), ao absorver os coeficientes de escala desconhecidos em parâmetros de localização e de escala da maneira usual, e levando a produtos através de blocos. Portanto, quando $\xi \neq 0$,

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{m} \left(\exp\left\{ -\left[1 + \xi \left(\frac{z_i^{(r_i)} - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\} \times \prod_{k=1}^{r_i} \sigma^{-1} \left[1 + \xi \left(\frac{z_i^{(k)} - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi} - 1} \right), \tag{3.26}$$

fornecida $1 + \xi \left(\frac{z_i^{(k)} - \mu}{\sigma} \right) > 0$, $k = 1, \ldots, r_i$, $i = 1, \ldots, m$; caso contrário, a verosimilhança é zero. Quando $\xi = 0$,

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{m} \left(\exp\left\{ -\exp\left[-\left(\frac{z^{(r_i)} - \mu}{\sigma} \right) \right] \right\} \times \prod_{k=1}^{r_i} \sigma^{-1} \exp\left[-\left(\frac{z_i^{(k)} - \mu}{\sigma} \right) \right] \right). \tag{3.27}$$

A verosimilhança (3.26) e (3.27) ou, mais frequentemente, a correspondente log-verosimilhança, pode ser maximizada numericamente para obter estimativas de MV. A teoria da verosimilhança assintótica padrão também fornece erros padrão e IC aproximados. No caso especial de $r_i = 1$ para cada i, a função de verosimilhança reduz-se à verosimilhança do modelo GEV dos blocos de máximos. De modo geral, através do modelo estatístico das r maiores observações obtém-se uma verosimilhança cujos parâmetros correspondem aos da distribuição GEV dos blocos de máximos, mas com mais quantidade de dados extremos observados incorporados. Portanto, em relação a uma análise de blocos de máximo padrão, a interpretação dos parâmetros é inalterada, mas a precisão deve ser melhorada, devido à inclusão de informações extras.

3.4 Modelos com Limiar

3.4.1 Introdução

Seja $X_1, X_2, \ldots$ uma sucessão de v.a.'s i.i.d., tendo como f.d. marginal F. É natural considerar como eventos extremos aqueles de X_i, que excedem algum limiar alto u. Denotando um termo arbitrário na sucessão X_i por X, segue-se que uma descrição do comportamento estocástico de eventos extremos é dada pela probabilidade condicional

$$\Pr\{X > u + y | X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \; y > 0, \tag{3.28}$$

se a distribuição principal F fosse conhecida, a distribuição de ultrapassagens do limiar em (3.28) também seria conhecida. Uma vez que, em aplicações práticas, este não é o caso, são procuradas aproximações que são amplamente aplicáveis para valores elevados do limiar. Isto é paralelo ao uso do modelo GEV, como uma aproximação da distribuição dos máximos das sucessões longas, quando a população principal é desconhecida.

3.4.2 Caracterização do Modelo Assintótico

3.4.2.1 Distribuição de Pareto Generalizada

O resultado principal está contido no seguinte teorema.

Teorema 11. *Seja $X_1, X_2, \ldots$ uma sucessão de v.a.'s independentes com a f.d. comum F e seja*

$$M_n = \max\{X_1, \ldots, X_n\}.$$

Denotando um termo arbitrário na sucessão X_i por X, e supondo que F satisfaz o Corolário 1, de modo que, para n grandes,

$$\Pr\{M_n \leq z\} \approx G(z),$$

onde

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\},$$

para alguns, $\mu, \sigma > 0$ e ξ. Então, para u suficientemente grande, a f.d. de $(X-u)$, condicional em $X > u$, é aproximadamente

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}, \tag{3.29}$$

definido em $\left\{y: y > 0 \ \ e \ \ \left(1 + \frac{\xi y}{\tilde{\sigma}}\right) > 0\right\}$, onde

$$\tilde{\sigma} = \sigma + \xi(u - \mu). \tag{3.30}$$

□

A família de distribuições definida pela (3.29) é chamada **Família Generalizada de Pareto (GP)**. O Teorema 11 implica que, se os blocos de máximos tiverem uma distribuição G aproximada, então os excessos de um limiar têm uma distribuição aproximada dentro da família GP. Além disso, os parâmetros desta distribuição dos limiares dos excessos são unicamente determinados por aqueles da distribuição GEV associados aos blocos de máximos. Em particular, o parâmetro ξ em (3.29) é igual ao da distribuição GEV correspondente. Escolhendo um parâmetro diferente, mas igualmente grande, o bloco de tamanho n afetaria os valores dos parâmetros GEV, mas não os da distribuição GP correspondente dos limiares dos excessos: ξ é invariante quanto ao tamanho do bloco, enquanto o cálculo de $\tilde{\sigma}$ em (3.30) não é perturbado pelas mudanças em μ e em σ que são auto-compensadoras.

A dualidade entre as famílias GEV e GP significa que o parâmetro de forma ξ é dominante na determinação do comportamento qualitativo da distribuição GP, assim como, para a distribuição GEV. Se $\xi < 0$, a distribuição de excessos tem um limiar superior de $u - \frac{\tilde{\sigma}}{\xi}$; se $\xi > 0$, a distribuição não tem limiar superior. Também poderá ser ilimitada se $\xi = 0$, que deve ser novamente interpretado tendo em conta o limiar $\xi \to 0$ em (3.29), levando a

$$H(y) = 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right), \; y > 0, \tag{3.31}$$

correspondendo a uma distribuição exponencial com o parâmetro $\frac{1}{\tilde{\sigma}}$.

3.4.2.2 Justificação do esboço do Modelo GP

Aqui apresenta-se uma pequena prova do Teorema 11, um argumento mais detalhado é dado em Leadbetter, Lindgren e Rootzen (1983).

Tendo X a f.d. F pela suposição do Teorema 8, para n suficientemente grande,

$$F^n(z) \approx \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}$$

para alguns parâmetros μ, $\sigma > 0$ e ξ. Consequentemente,

$$n\log F(z) \approx \left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}. \tag{3.32}$$

Mas para grandes valores de z, a expansão em série de Taylor implica que

$$\log F(z) \approx \; -\{1 - F(z)\}.$$

Substituindo em (3.32) obtém-se

$$n\left(-\{1 - F(z)\}\right) \approx -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}},$$

da qual se obtém

$$1 - F(u) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}},$$

para u grande. Da mesma forma, para $y > 0$,

$$1 - F(u+y) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u+y-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}. \tag{3.33}$$

Por isso,

$$\Pr\{X > u + y | X > u\} \approx \frac{n^{-1}\left[1 + \frac{\xi(u+y-\mu)}{\sigma}\right]^{-\frac{1}{\xi}}}{n^{-1}\left[1 + \frac{\xi(u-\mu)}{\sigma}\right]^{-\frac{1}{\xi}}} = \left[1 + \frac{\frac{\xi(u+y-\mu)}{\sigma}}{1 + \frac{\xi(u-\mu)}{\sigma}}\right]^{-\frac{1}{\xi}} = \left[1 + \frac{\xi y}{\tilde{\sigma}}\right]^{-\frac{1}{\xi}}, \tag{3.34}$$

onde,

$$\tilde{\sigma} = \sigma + \xi(u - \mu),$$

como requerido.

3.4.3 Modelação dos limiares dos excessos

3.4.3.1 Seleção do limiar

O Teorema 11 sugere a seguinte estrutura para modelação de valores extremos. Os dados em bruto consistem numa sucessão de medidas $x_1,\ldots,x_n$. Eventos extremos são identificados por um limiar alto u, para o qual as excedências são $\{x_i : x_i > u\}$. Rotulando estas excedências por $x_{(1)},\ldots,x_{(k)}$ e definindo o limiar dos excessos por $y_j = x_{(j)} - u$, sendo que $j = 1,\ldots,k$. Por este teorema, o y_j pode ser considerado como realizações independentes de uma v.a. cuja distribuição pode ser aproximada por um membro da família GP. A inferência consiste em ajustar a família GP ao limiar de excedências observado, seguido da verificação e extrapolação do modelo.

Esta abordagem contrasta com a abordagem dos blocos de máximos através da caracterização de uma observação como extrema se exceder a um limiar alto. Mas a questão da escolha do limiar é análoga à escolha do tamanho do bloco na abordagem dos blocos de máximos, implicando um equilíbrio entre viés e variância. Neste caso, um limiar muito baixo é suscetível de não respeitar a base assintótica do modelo, levando ao enviesamento; se o limiar for muito alto irá gerar alguns excessos com os quais o modelo pode ser estimado, o que levará a uma alta variância. A prática padrão é adotar como limiar o mais baixo possível, que levará, em princípio, a um modelo com limiar que fornece uma aproximação razoável. Existem dois métodos disponíveis para este fim: um é uma técnica exploratória realizada antes da estimação do modelo; a outra é uma avaliação da estabilidade das estimativas dos parâmetros, baseada na adaptação de modelos numa gama de diferentes limiares.

Mais detalhadamente, o primeiro método é baseado na média da distribuição GP. Se Y tem uma distribuição GP com parâmetros σ e ξ, então

$$E(Y) = \frac{\sigma}{1-\xi}, \tag{3.35}$$

fornecido $\xi < 1$. Quando $\xi \geq 1$ a média é infinita. Agora, supondo que a distribuição GP é válida como modelo para os excessos de um limiar u_0 gerado por uma série $X_1,\ldots,X_n$, da qual um termo arbitrário é denotado por X. Pela (3.35),

$$E(X - u_0 | X > u_0) = \frac{\sigma_{u_0}}{1-\xi},$$

fornecido $\xi < 1$, onde se denota σ_{u_0} como parâmetro de escala correspondente aos excessos do limiar u_0. Mas se a distribuição GP é válida para os excessos do limiar u_0, deve igualmente ser válida para todos os limiares $u > u_0$, sujeita à mudança do parâmetro de escala apropriada para σ_u. Portanto, para $u > u_0$,

$$E(X - u | X > u) = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u_0} + \xi u}{1-\xi} \tag{3.36}$$

em virtude de (3.30). Então, para $u > u_0$, $E(X - u \mid X > u)$ é uma função linear de u. Além disso, $E(X - u \mid X > u)$ é simplesmente a média dos excessos do limiar u, para o qual a

média da amostra dos excessos do limiar u fornece uma estimativa empírica. De acordo com (3.36), estas estimativas são esperadas mudar linearmente com u, em níveis de u para os quais o modelo da GP é apropriado. Isto leva ao seguinte procedimento. O lugar geométrico dos pontos

$$\left\{\left(u, \frac{1}{n_u}\sum_{i=1}^{n_u}(x_{(i)}-u)\right): u < x_{max}\right\},$$

onde $x_{(1)},\ldots,x_{(n_u)}$ consistem nas n_u observações que excedem u e x_{max} é o maior dos X_i, é denominado por **gráfico de vida residual média (GVRM)**. Acima de um limiar u_0, em que a distribuição GP fornece uma aproximação válida para a distribuição excesso, o GVRM deve ser, aproximadamente, linear em u. Os IC podem ser adicionados ao gráfico com base na normalidade aproximada das médias de amostragem. A interpretação de um GVRM nem sempre é simples na prática.

O segundo procedimento, para seleção de limiares, é estimar o modelo numa gama de limiares. Acima de um nível u_0, em que a motivação assintótica para a distribuição GP é válida, as estimativas do parâmetro da forma, ξ, devem ser, aproximadamente, constantes, enquanto as estimativas de σ_u devem ser lineares em u, devido a (3.36).

3.4.3.2 Estimação de Parâmetros

Tendo determinado um limiar, os parâmetros da distribuição GP podem ser estimados pela MV. Supondo que os valores $y_1, \ldots, y_k$ são os k excessos de um limiar u. Para $\xi \neq 0$ a log-verosimilhança é derivada a partir de (3.29) como

$$\ell(\sigma,\xi) = -k\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\log\left(1+\frac{\xi y_i}{\sigma}\right), \tag{3.37}$$

dado $\left(1+\frac{\xi y_i}{\sigma}\right) > 0$ para $i = 1,\ldots,k$; de outro modo, $\ell(\sigma,\xi) = -\infty$. No caso de $\xi = 0$ a log-verosimilhança é obtida da (3.31) como

$$\ell(\sigma) = -k\log\sigma - \sigma^{-1}\sum_{i=1}^{k} y_i.$$

A maximização analítica da log-verosimilhança não é possível, por isso, são novamente necessárias técnicas numéricas, é preciso cuidado para evitar instabilidades numéricas quando $\xi \approx 0$ em (3.37), e é necessário assegurar que o algoritmo não falhe, devido à avaliação feita fora do espaço de parâmetros permitido. Os erros padrão e os IC para a distribuição GP são obtidos da forma habitual da teoria da verosimilhança padrão.

3.4.3.3 Níveis de Retorno

Como já referido, é geralmente mais conveniente interpretar modelos de valores extremos em termos de quantis ou NR, em vez de valores de parâmetros individuais. Por isso,

assumindo que uma distribuição GP com os parâmetros σ e ξ é um modelo adequado para excedências de um limiar u por uma variável X. Ou seja, para $x > u$,

$$\Pr\{X > x | X > u\} = \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\xi}}.$$

Segue que

$$\Pr\{X > x\} = \zeta_u \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\xi}}, \tag{3.38}$$

onde $\zeta_u = \Pr\{X > u\}$. Assim, o nível x_m que é excedido, em média, uma vez a cada m observações, é a solução de

$$\zeta_u \left[1 + \xi\left(\frac{x_m - u}{\sigma}\right)\right]^{-\frac{1}{\xi}} = \frac{1}{m}. \tag{3.39}$$

Reorganizando fica,

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right], \tag{3.40}$$

desde que m seja suficientemente grande para garantir que $x_m > u$. Isto tudo assume que $\xi \neq 0$. Se $\xi = 0$, fazendo o mesmo com (3.31) leva a

$$x_m = u + \sigma \log(m\zeta_u)\,, \tag{3.41}$$

novamente, desde que m seja suficientemente grande.

Por construção, x_m é o **nível de retorno da observação *m***. A partir da (3.40) e da (3.41), ao se fazer um gráfico de x_m contra m numa escala logarítmica, produz-se as mesmas características qualitativas como nos gráficos de NR baseados no modelo GEV: linearidade se $\xi = 0$; concavidade se $\xi > 0$; convexidade se $\xi < 0$. Para apresentar, é mais conveniente mostrar os NR numa escala anual, de modo que o NR do ano N seja o nível esperado para ser excedido uma vez a cada N anos. Se existem n_y observações por ano, isto corresponde ao NR da observação m, onde $m = N \times n_y$. Assim, o NR do ano N é definido por

$$z_N = u + \frac{\sigma}{\xi}\left[\left(N n_y \zeta_u\right)^{\xi} - 1\right],$$

a menos que $\xi = 0$, nesse caso

$$z_N = u + \sigma \log(N n_y \zeta_u)\,.$$

A estimação dos NR requer a substituição dos valores dos parâmetros pelas suas estimativas. Para σ e ξ isto corresponde à substituição pelas estimativas de MV correspondentes, e a estimativa de ζ_u, ou seja, a probabilidade de uma observação individual exceder o limiar u, também é necessária. Terá um estimador natural de

$$\hat{\zeta}_u = \frac{k}{n},$$

a proporção da amostra de pontos que excede u. Uma vez que, o número de excedências de u seguem uma distribuição binomial $Bin(n,\ \zeta_u)$, $\hat{\zeta}_u$, é também um estimador de MV de ζ_u.

Os erros padrão ou os IC para x_m podem ser derivados pelo método delta, mas a incerteza na estimativa de ζ_u também deve ser incluída no cálculo. A partir das propriedades padrão da distribuição binomial $Var(\hat{\zeta}_u) \approx \frac{\hat{\zeta}_u(1-\hat{\zeta}_u)}{n}$, então a matriz completa de variância-covariância para $(\hat{\zeta}_u, \hat{\sigma}, \hat{\xi})$ é aproximadamente

$$V = \begin{bmatrix} \frac{\hat{\zeta}_u(1-\hat{\zeta}_u)}{n} & 0 & 0 \\ 0 & v_{1,1} & v_{1,2} \\ 0 & v_{2,1} & v_{2,2} \end{bmatrix},$$

onde $v_{i,j}$ denota o termo (i,j) da matriz de variância-covariância de $\hat{\sigma}$ e $\hat{\xi}$. Assim, pelo método delta,

$$Var(\hat{x}_m) \approx \nabla {x_m}^T V \nabla x_m, \tag{3.42}$$

onde

$$\nabla {x_m}^T = \left[\frac{\partial x_m}{\partial \zeta_u}, \frac{\partial x_m}{\partial \sigma}, \frac{\partial x_m}{\partial \xi} \right] =$$

$$= \left[\sigma m^{\xi} {\zeta_u}^{\xi-1}, \xi^{-1} \left\{ (m\zeta_u)^{\xi} - 1 \right\}, -\sigma \xi^{-2} \left\{ (m\zeta_u)^{\xi} - 1 \right\} + \sigma \xi^{-1} (m\zeta_u)^{\xi} \log(m\zeta_u) \right],$$

avaliado em $(\hat{\zeta}_u, \hat{\sigma}, \hat{\xi})$.

Como nos modelos anteriores, as melhores estimativas de precisão para os parâmetros e os NR são obtidos a partir do perfil apropriado de verosimilhança. Para σ ou ξ isto é simples, para os NR, é requerida uma reparametrização. É mais simples ignorar a incerteza em ζ_u, que é geralmente pequena em relação à dos outros parâmetros. A partir de (3.40) e (3.41)

$$\sigma = \begin{cases} \frac{(x_m - u)\xi}{(m\zeta_u)^{\xi} - 1}, & se\ \xi \neq 0; \\ \frac{x_m - u}{\log(m\zeta_u)}, & se\ \xi = 0. \end{cases}$$

Com x_m fixo, a substituição em (3.37) leva a uma verosimilhança de um parâmetro que pode ser maximizada em relação a ξ. Como função do x_m, este é o perfil de log-verosimilhança para o NR da observação m.

3.4.3.4 Escolha do limiar revista

Como foi mencionado, os GVRM podem ser difíceis de interpretar como um método de seleção de limiares. Uma técnica complementar é ajustar a distribuição GP numa gama de limiares e procurar a estabilidade das estimativas dos parâmetros. O argumento é o seguinte.

Pelo Teorema 11, se uma distribuição GP for um modelo razoável para excessos de um limiar u_0, então os excessos de um limiar superior u também devem seguir uma distribuição GP. Os parâmetros de forma, das duas distribuições, são idênticos. No entanto, denotando por σ_u, o valor do parâmetro de escala da distribuição GP, para um limiar de $u > u_0$, segue-se de (3.30) que

$$\sigma_u = \sigma_{u_0} + \xi (u - u_0), \tag{3.43}$$

de modo que o parâmetro de escala mude com u a menos que $\xi = 0$. Esta dificuldade pode ser reparada, ao modificar o parâmetro de escala da distribuição GP do seguinte modo,

$$\sigma^* = \sigma_u - \xi u,$$

que é constante em relação a u em virtude de (3.43). Consequentemente, as estimativas de ambos σ^* e ξ devem ser constantes acima de u_0, se u_0 é um limiar válido de excessos para acompanhar a distribuição GP. A variabilidade da amostra significa que as estimativas destas quantidades não serão exatamente constantes, mas devem ser estáveis após a permissão para os seus erros de amostragem.

Este argumento sugere o gráfico de $\hat{\sigma}^*$ e $\hat{\xi}$ contra u, junto com os IC para cada uma dessas quantidades, e selecionando u_0 como o menor valor de u, para o qual as estimativas permanecem quase constantes. Os IC de $\hat{\xi}$ são obtidos imediatamente a partir da matriz variância-covariância V. Os IC para $\hat{\sigma}^*$ requerem o método delta, usando

$$Var(\hat{\sigma}^*) \approx \nabla\sigma^{*T} V \nabla\sigma^*,$$

onde

$$\nabla\sigma^{*T} = \left[\frac{\partial\sigma^*}{\partial\sigma_u}, \frac{\partial\sigma^*}{\partial\xi}\right] = [1, -u].$$

3.4.3.5 Verificação do Modelo

Gráficos de probabilidade, de quantis, de NR e de densidade são todos úteis para avaliar a qualidade de um ajuste do modelo GP. Assumindo um limiar u, os limiares de excessos $y_{(1)} \leq \cdots \leq y_{(k)}$ e um modelo estimado $\hat{H}$, o gráfico de probabilidade consiste nos pares $\left\{\left(\frac{i}{k+1}, \hat{H}(y_{(i)})\right);\ i = 1, \ldots, k\right\}$, onde

$$\hat{H}(y) = 1 - \left(1 + \frac{\hat{\xi} y}{\hat{\sigma}}\right)^{-\frac{1}{\hat{\xi}}},$$

fornecido $\hat{\xi} \neq 0$. Se $\hat{\xi} = 0$ o gráfico é construído usando (3.31) no lugar de (3.29). Novamente assumindo $\hat{\xi} \neq 0$, o gráfico de quantis consiste nos pares

$$\left\{\left(\hat{H}^{-1}\left(\frac{i}{k+1}\right), y_{(i)}\right);\ i = 1, \ldots, k\right\},$$

onde

$$\hat{H}^{-1}(y) = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[y^{-\hat{\xi}} - 1\right].$$

Se o modelo GP for razoável para modelar excessos de u, então ambos os gráficos de probabilidade e de quantis devem consistir em pontos que são aproximadamente lineares.

Um gráfico de NR, consiste no lugar geométrico dos pontos $\{(m,\ \hat{x}_m)\}$ para grandes valores de m, onde $\hat{x}_m$ é o NR estimado da observação m:

$$\hat{x}_m = u + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(m\hat{\zeta}_u\right)^{\hat{\xi}} - 1\right],$$

novamente modificado se $\hat{\xi} = 0$. Tal como acontece com o gráfico de NR do modelo GEV, é normal traçar a curva de NR numa escala logarítmica para enfatizar o efeito de extrapolação e também para adicionar limites de confiança e estimativas empíricas dos NR.

Finalmente, a função densidade do modelo GP ajustado pode ser comparado com um histograma das excedências dos limiares.

CAPÍTULO 4

Aplicação de Modelos de Valores Extremos e análise dos resultados

4.1 Introdução

Nesta secção, serão aplicados cada um dos modelos descritos em detalhe no capítulo 3, aos dados do tráfego diário da Ponte 25 de Abril e tirar-se-ão algumas conclusões sobre a aplicabilidade dos modelos a estes dados.

Vai-se fazer uma análise preliminar aos dados essencialmente gráfica. Esta permitirá ter uma ideia do comportamento da cauda direita da distribuição associada aos dados referentes aos máximos anuais de veículos na Ponte 25 de Abril. A amostra, como já foi indicada, é composta pelo número de veículos que passaram diariamente na Ponte 25 de Abril, desde 01 de janeiro de 2010 a 31 de dezembro de 2018.

Tal como anteriormente foi explicado no capítulo 3, para se modelar os extremos de uma série de observações independentes, os dados são agrupados em blocos de observações de cumprimento n, gerando uma série de máximos de blocos. A dimensão dos blocos será escolhida para corresponder a um período de tempo de um ano, portanto, n será o número de observações num ano e os máximos dos blocos são máximos anuais, já que deste modo a sazonalidade não irá afetar a análise dos dados. Assumindo que este valor de n seja suficientemente grande, os argumentos assintóticos levam a um modelo que descreve as variações nos máximos anuais de um ano para o outro e que podem ser ajustadas aos máximos anuais observados.

No entanto, em qualquer ano em particular, podem ter ocorrido eventos extremos adicionais e é possível que sejam mais extremos do que o máximo de outros anos. Já que tais dados não são o máximo anual no ano em que surgiram, estes vão ser excluídos de uma parte desta análise, contudo, na aplicação do Modelo GEV Multivariado e no Modelo estatístico das r maiores observações serão tidos em consideração mais valores para além do máximo anual. No caso em que só se considera o máximo anual, como se têm os dados

diários de 9 anos a amostra terá uma dimensão igual a $m = 9$ observações e esta está representada na figura 4.1.

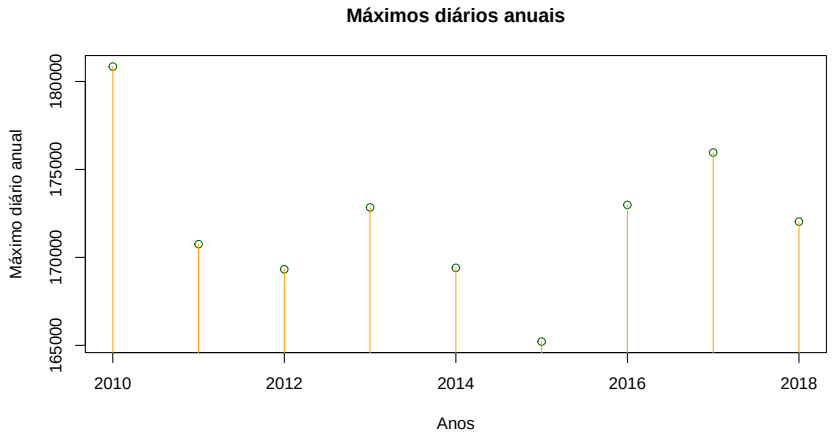


Figura 4.1: Máximos diários anuais do tráfego na Ponte 25 de Abril (2010-2018)

No gráfico de autocorrelação parcial (figura 4.2), apresentado de seguida, pode-se observar que os valores são fracamente correlacionados entre si, portanto, é possível a existência de independência nos dados.

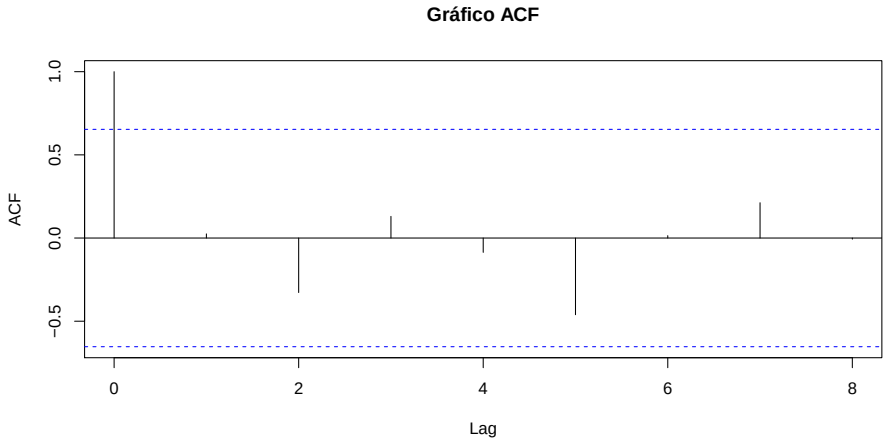


Figura 4.2: Gráfico da Autocorrelação Parcial

As características amostrais, como a mediana, a média, os quartis e os extremos são:

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
165212	169406	172030	172150	172982	180846

Figura 4.3: Características Amostrais

Observa-se através destes valores que o número de veículos varia entre 165212 e

180846. Verifica-se que os dados são negativamente assimétricos pelo boxplot representado na figura 4.4, tira-se essa conclusão pelo risco da mediana que se encontra mais próximo do 3º Quartil. Também se tem um *outlier* representado, ou seja, um ponto fora das "linhas" desenhadas, no ponto máximo da distribuição, ou seja, em 180846.

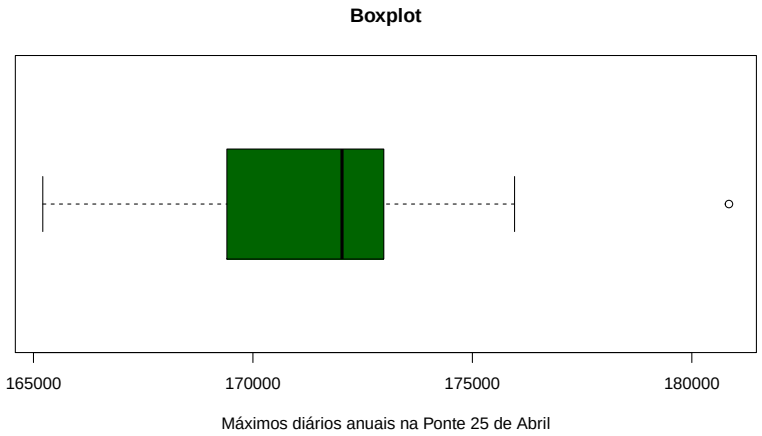


Figura 4.4: Boxplot dos máximos diários anuais na Ponte 25 de Abril (2010-2018)

De seguida para se efetuarem as aplicações dos Modelos de Valores Extremos foram tidos em conta, para além do livro Coles (2001), a dissertação de mestrado Rosário (2013) e "GITHUB" (2009). As aplicações efetuadas ao tráfego da Ponte 25 de Abril nomeadamente: ao Modelo GEV os blocos dos máximos anuais; ao Modelo GEV Multivariado as 3, 5 e 10 maiores observações anuais; e ao Modelo GP os valores do tráfego acima do limiar u igual a 165212, 156297 e 161734 ao Modelo GP.

4.2 Modelo GEV

Como mencionado na secção 3.3.1.3 para se modelarem valores extremos de uma série de observações independentes $X_1, X_2, \ldots$ os dados juntam-se em sucessões de observações de comprimento n, gerando uma série de blocos de máximos, $M_{n,1}, \ldots, M_{n,m}$ (para os quais a distribuição GEV poderá ser montada). Escolheu-se o comprimento de um ano, portanto, n será o número de observações num ano e os máximos dos blocos serão máximos anuais.

Como já mencionado anteriormente os máximos anuais são os seguintes:

Blocos	Data do valor Máximo	Valor Máximo
$M_{1,1}$	02/07/2010	180 846
$M_{1,2}$	15/07/2011	170 750
$M_{1,3}$	06/07/2012	169 322
$M_{1,4}$	28/06/2013	172 842
$M_{1,5}$	11/07/2014	169 406
$M_{1,6}$	26/06/2015	165 212
$M_{1,7}$	15/07/2016	172 982
$M_{1,8}$	14/07/2017	175 961
$M_{1,9}$	31/08/2018	172 030

Tabela 4.1: Blocos de máximos, valores dos máximos anuais e respetivas datas

Os blocos de máximos serão denotados por $Z_1, \ldots, Z_m$, neste caso, com $m = 9$.

4.2.0.1 Estimação por Máxima Verosimilhança

Os dados são modelados como observações independentes da distribuição GEV. Quanto à maximização da log-verosimilhança do GEV, obtida através da função *gev.fit()* do pacote *ismev* (Heffernan & Stephenson, 2018) que será utilizado ao longo das três secções seguintes (sendo esta a primeira), para os dados referidos, obteve-se o seguinte *output*:

R code 4.1: *Output do gev.fit()*

```
$conv
[1] 0

$nllh
[1] 87.59868

$mle
[1]  1.701567e+05  3.778887e+03 -1.324375e-01

$se
[1] 1403.6701498 1027.8844808    0.2397786
```

Pode-se verificar que:

$$(\hat{\mu}, \hat{\sigma}, \hat{\xi}) = (170156.7, 3778.887, -0.1324375),$$

para a qual a log-verosimilhança é −87.59868. A matriz da variância-covariância aproximada das estimativas dos parâmetros é:

$$M = \begin{bmatrix} 1970289.8894 & 183742.0436 & -109.8018 \\ 183742.0436 & 1056546.5059 & -113.8859 \\ -109.8018 & -113.8859 & 0.05749378 \end{bmatrix}$$

Tendo em consideração os resultados obtidos por Smith (1985), e já que $\hat{\xi} > -0.5$, pode-se afirmar que os estimadores da MV são regulares, ou seja, têm as propriedades assintóticas usuais.

A diagonal principal da matriz representada corresponde aos valores das variâncias dos parâmetros individuais de (μ, σ, ξ). Calculando as respetivas raízes quadradas, obtêm-se os erros padrão que são 1403.670,1027.884 e 0.2397786 para $\hat{\mu}$, $\hat{\sigma}$ e $\hat{\xi}$ respetivamente. Podem-se calcular os IC de 95%, aproximadamente, para cada parâmetro, combinando as estimativas obtidas e os erros padrão:

Parâmetro	IC de 95%
μ	[167405.5,172907.8]
σ	[1764.234,5793.541]
ξ	[−0.6024036,0.3375286]

Tabela 4.2: Valores dos IC dos parâmetros estimados.

Como se pode ver pelos valores dos ICs dos parâmetros estimados, o IC do parâmetro ξ contém zero, logo a Distribuição Gumbel poderá ser a distribuição mais precisa da família GEV para estes dados. Para analisar esta situação será efetuado o ajustamento dos dados à distribuição Gumbel na secção 4.2.0.5.

4.2.0.2 Verificação do Modelo

Para se visualizar melhor a extrapolação do modelo GEV, tem-se a figura 4.5. Para isso, utilizou-se a função *gev.diag()*, que para modelos estacionários produz quatro gráficos diagnóstico (usando o *output* da função *gev.fit()*).

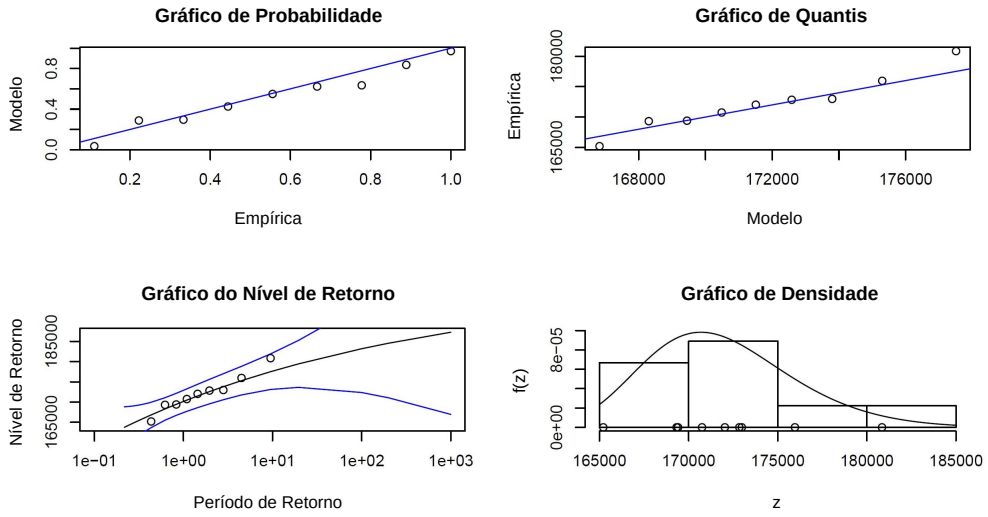


Figura 4.5: Gráficos diagnóstico para o Modelo GEV ajustado aos dados do tráfego da Ponte 25 de Abril

Através do gráfico de probabilidade, que é uma comparação entre funções de distribuição empírica e ajustada, quando os pontos ficam perto da diagonal unidade é sinal de que o modelo GEV está bem ajustado. Se há desvios substanciais da linearidade estes são indicativos de alguma falha no modelo. Como aparentemente não acontece pode-se afirmar o contrário.

No gráfico de quantis se estiverem representadas saídas de linearidade, tal como no caso anterior, estas indicariam falha do modelo. Apesar de haver um ponto um pouco afastado da diagonal, todos os outros apresentam uma certa linearidade. Tem-se também o gráfico baseado na função de densidade, ou seja, está representada uma comparação da função de densidade de probabilidade (de um modelo ajustado) com um histograma dos dados.

Quanto ao gráfico do NR tem-se a representação de um gráfico do nível que se espera que seja excedido pelo processo uma vez em cada p anos (nível de retorno z_p) contra o (logaritmo do) período de retorno p. O gráfico do NR é particularmente relevante para interpretar modelos de valor extremo. A cauda da distribuição é comprimida de tal modo que as estimativas do NR são exibidas para longos períodos de retorno. A linha preta representa a estimativa da MV dos parâmetros da distribuição GEV ajustada aos dados do tráfego da Ponte 25 de Abril. As linhas azuis são IC de aproximadamente 95%. Já os pontos são os níveis de retorno empíricos e ajudam na validação do modelo, portanto, neste caso, existem 9 pontos no conjunto de dados, o maior ponto corresponde ao quantil empírico do ano 9. Tendo em conta a observação do gráfico de NR da figura 4.5 pode-se dizer que o modelo está bem ajustado, já que os pontos se encontram entre as linhas de confiança.

Os gráficos da figura 4.5 que têm como base uma comparação entre modelos base

e estimativas empíricas da f.d., estão razoavelmente de acordo quanto à adequação do modelo GEV ajustado aos dados referentes aos máximos anuais do tráfego na Ponte 25 de Abril.

4.2.0.3 Inferência para níveis de retorno

As estimativas para os NR são obtidas pela substituição dos valores nas equações (3.22) e (3.23). Para se calcularem os IC de 95% será calculada a variância do NR, pelo método delta. Vão ser calculados quatro NR para: 5, 10, 50 e 100 anos. Feitas as respetivas substituições nas equações e calculadas as mesmas, foram obtidos os seguintes valores:

N p	$y_p = -\log(1-p)$	Nível de retorno $(\hat{z}_p)$	IC de 95%
5 anos $p = 0.2$	$y_{0.2} = -\log(0.8)$	175297	[171763,178831]
10 anos $p = 0.1$	$y_{0.1} = -\log(0.9)$	177510	[173115,181906]
50 anos $p = 0.02$	$y_{0.02} = -\log(0.98)$	181671	[173267,190076]
100 anos $p = 0.01$	$y_{0.01} = -\log(0.99)$	183174	[172367,193982]

Tabela 4.3: Valores obtidos para diferentes anos de NR para o modelo GEV

Tendo em conta a tabela 4.3 espera-se que, em média, num ano em cada 5, 10, 50 ou 100 anos, haja um dia em que o número de veículos que atravessa a Ponte 25 de Abril seja superior a 175297, 177510, 181671 e 183174, respetivamente.

Como, neste caso, $\hat{\xi} < 0$ também é possível fazer inferências sobre o limite superior do suporte da distribuição que é efetivamente o 'período inferior de retorno da observação', ou seja, calcula-se $\hat{z}_p$ com $p = 0$.

A estimativa da MV é $\hat{z}_0 = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}$ e (3.20) é válida com $\nabla z_0{}^T = \left[1, -\xi^{-1}, \sigma\xi^{-2}\right]$ avaliado em $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$. Portanto, para este caso em concreto, depois de efetuadas as respetivas substituições e cálculos, tem-se $\hat{z}_0 = 198690$ e o respetivo IC de, aproximadamente, 95% é $[104368, 293012]$. Como se pode observar o valor de $\hat{z}$ para $p = 0$ é o maior valor comparando com os outros resultados, como seria de esperar, e em relação ao IC é também o de maior amplitude.

4.2.0.4 Perfil da Verosimilhança

Para se obter o perfil da verosimilhança vai-se usar a função *gev.prof()* do pacote *ismev* (Heffernan & Stephenson, 2018). Esta função permite o cálculo do perfil log-verosimilhança para o parâmetro de forma, ξ, e para diferentes anos de NRs, para modelos

GEV.

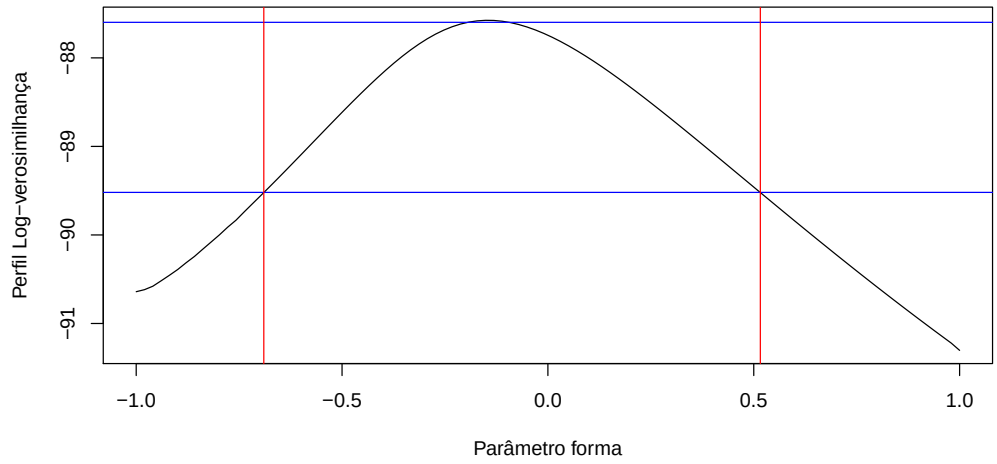
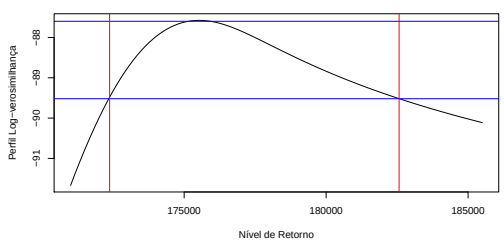
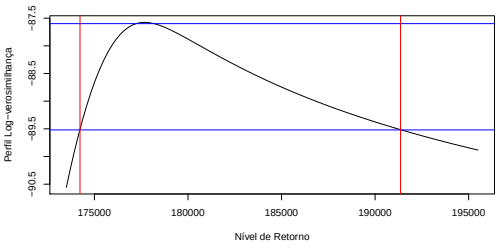


Figura 4.6: Perfil da log-verosimilhança para ξ para os máximos anuais do tráfego da Ponte 25 de Abril

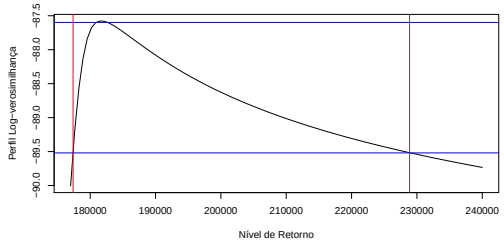
A figura 4.6 mostra o gráfico do perfil da log-verosimilhança para ξ no tráfego da Ponte 25 de Abril cujos valores do IC de, aproximadamente, 95%, obtidos através do mesmo, são $[-0.6900517, 0.516]$. Têm-se aqui ilustrados os quatro gráficos do perfil log-verosimilhança para os diferentes NR:



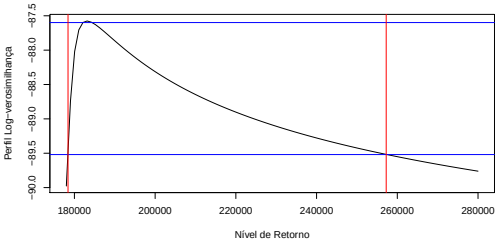
(a) NR de 5 anos



(b) NR de 10 anos



(c) NR de 50 anos



(d) NR de 100 anos

Figura 4.7: Perfil da log-verosimilhança para diferentes anos de NR no tráfego da Ponte 25 de Abril

Um IC de, aproximadamente, 95% para o NR de 5 anos é obtido a partir do perfil da log-verosimilhança como [172374, 182571]; para 10 anos é [174230, 191357]; para 50 anos é [177380, 228875]; para 100 anos é [178418, 257250].

4.2.0.5 Distribuição de Gumbel

Segundo o que já foi referido, em relação às distribuições de valores extremos, quando se obtém o parâmetro de forma menor que zero, em princípio, significaria que a distribuição em causa seria do tipo Weibull. No entanto, vai-se fazer a substituição da família GEV pela família Gumbel que corresponde a $\xi = 0$, já que na secção 4.2.0.1 se verificou que o IC do parâmetro ξ contém o valor zero, para isso usa-se a função *gum.fit()* do mesmo pacote para a obtenção da estimação dos parâmetros. Para o caso aqui estudado obtiveram-se os seguintes resultados:

R code 4.2: *Output do gum.fit()*

```
$conv
[1] 0

$nllh
[1] 87.7446

$mle
[1] 170160.479   3665.779

$se
[1] 1293.3233  915.7362
```

Quanto aos gráficos de diagnóstico utiliza-se a função *gum.diag()*, com o *output* anterior, e resultaram os gráficos seguintes:

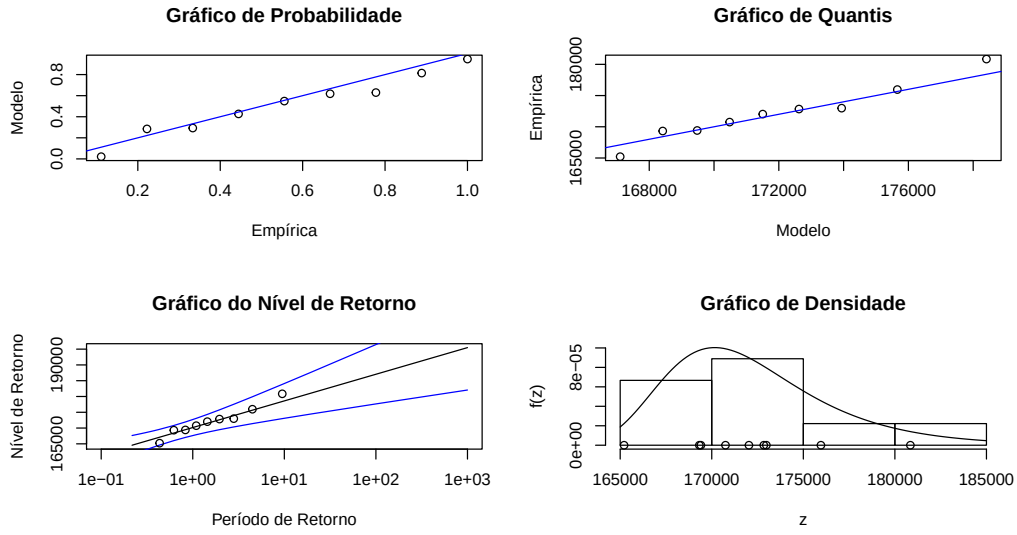


Figura 4.8: Gráficos diagnóstico para o ajuste do Modelo Gumbel aos máximos anuais do tráfego da Ponte 25 de Abril

Neste caso, a MV corresponde à maximização da (3.21) e depois dos cálculos correspondentes obtêm-se os erros padrão e não só. Para os dados aqui estudados, como já se observou pelo *output*, obteve-se:

$$(\hat{\mu}, \hat{\sigma}) = (170160.479, 3665.779)$$

Já os erros padrão são 1293.3233 e 915.7362 para $\hat{\mu}$ e $\hat{\sigma}$, respetivamente, a log-verosimilhança é de −87.7446.

Tendo em consideração, os resultados obtidos, pode-se calcular a estatística do teste da razão de verosimilhança para a redução do modelo Gumbel:

$$D = 2\{87.7446 - 87.59868\} = 0.29184$$

Este valor é pequeno quando comparado com a distribuição χ^2_1, o que sugere que o modelo de Gumbel é adequado para estes dados. Já se tinha essa impressão e ao se visualizarem os gráficos diagnóstico na figura 4.8, que mostram que a qualidade do ajuste é comparável à do modelo GEV. Não é nenhuma surpresa, uma vez que os parâmetros estimados nestes dois modelos são tão semelhantes, o que também significa que (a curto prazo) a extrapolação do modelo com base em qualquer dos modelos leva a respostas bastante semelhantes.

Pode-se fazer uma comparação em relação aos IC. Os valores obtidos para os últimos parâmetros estimados são:

Parâmetro	IC de 95%
μ	[167625.566,172695.393]
σ	[1870.937,5460.622]

Tabela 4.4: Valores dos IC dos parâmetros estimados pelo modelo Gumbel

As estimativas para os NR são obtidas pela substituição dos valores nas equações (3.22) e (3.23), neste caso, na equação (3.22) será substituído o segundo ramo do sistema. Para se calcularem os IC de, aproximadamente, 95% será calculada, como no caso anterior, pelo método delta. Os cálculos foram efetuados para os mesmos anos de NR e obtiveram-se os seguintes valores:

N p	$y_p = -\log(1-p)$	Nível de retorno $(\hat{z}_p)$	IC de 95%
5 anos $p = 0.2$	$y_{0.2} = -\log(0.8)$	175659	[171405,179913]
10 anos $p = 0.1$	$y_{0.1} = -\log(0.9)$	178410	[172990,183829]
50 anos $p = 0.02$	$y_{0.02} = -\log(0.98)$	184464	[176280,192648]
100 anos $p = 0.01$	$y_{0.01} = -\log(0.99)$	187024	[177635,196413]

Tabela 4.5: Valores obtidos para diferentes anos de NR para o modelo Gumbel

Segundo a tabela 4.5 espera-se que, em média, num ano em cada 5, 10, 50 ou 100 anos, haja um dia em que o número de veículos que atravessa a Ponte 25 de Abril seja superior a 175659, 178410, 1844641 e 187024, respetivamente.

Posto isto, a maior diferença entre os dois modelos é em termos de precisão de estimação, ou seja, os parâmetros dos modelos têm estimativas com IC de menor amplitude no modelo Gumbel. Já em relação às estimativas dos IC para os NR de 5 e de 10 anos, os IC são de maior amplitude no modelo Gumbel e para os NR de 50 e de 100 anos são maiores no modelo GEV.

Para se optar por um dos modelos são de grande ajuda os gráficos diagnóstico. As estimativas das curvas do NR são bastante semelhantes, no entanto, os IC são mais amplos no modelo GEV, especialmente para períodos de retorno mais longos. Uma incerteza reduzida é sempre desejável, de modo que se o modelo Gumbel pudesse ser mais confiável, as suas inferências seriam preferidas. Sabe-se que o teorema de modelos extremos fornece suporte para se modelar os blocos de máximos com a família GEV, da qual a família Gumbel é um subconjunto. Realmente, verifica-se através dos dados que o modelo Gumbel é

plausível, por outro lado, isso não implica que os outros modelos não sejam.

De facto, a estimativa da MV dentro da família GEV não é da família Gumbel, portanto, a opção mais segura é aceitar que há incerteza sobre o valor do parâmetro forma e preferir a inferência baseada no modelo GEV.

4.3 Modelo GEV Multivariado

Na análise de valores extremos, existe uma dificuldade que é a quantidade limitada de dados para a estimativa do modelo. Nesta secção, vai-se utilizar uma caracterização geral, que é baseada no comportamento das estatísticas das r maiores observações dentro de um bloco, para valores de r pequenos.

Tendo em conta que $X_1, X_2, \ldots$ é uma sucessão de v.a.'s i.i.d. que representa, neste caso, os valores diários do tráfego da Ponte 25 de Abril e tem como objetivo caracterizar o comportamento do extremo X_i. Primeiro vai-se estender o resultado obtido na secção 3.3.1.3 para outras estatísticas de ordem extrema, definindo

$$M_n{}^{(k)} = k \text{ maior estatística ordinal de } \{X_1, \ldots, X_n\},$$

e identificando o comportamento do limite dessa variável, para k fixo, com $n \to \infty$.

Deste modo, como o objetivo é aplicar o modelo para cada bloco de um ano, vão se extrair os $k = 3$, $k = 5$ e $k = 10$ maiores valores diários de tráfego, obtendo-se um conjunto de 9 vetores aleatórios 3-dimensionais, 5-dimensionais e 10-dimensionais. Foram elegidos os valores $k = 3$ e $k = 5$, tendo em conta, os gráficos de probabilidade e de quantis I.3 em anexo, já a seleção do $k = 10$ foi principalmente com o objetivo de ter um k superior como termo de comparação. As observações estão representadas nas figuras 4.9, 4.10 e 4.11.

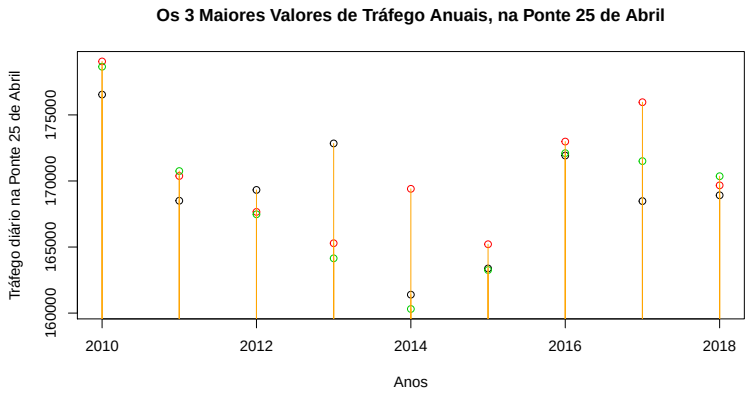


Figura 4.9: Os 3 maiores valores diários de tráfego, por ano, na Ponte 25 de Abril (2010-2018)

O teorema 9 implica que se a estatística das r maiores observações num bloco for normalizada exatamente da mesma maneira que o máximo, então a sua distribuição limite é dada por (3.23) cujos parâmetros correspondem aos parâmetros da distribuição

limite de GEV do bloco de máximos. Existe uma dificuldade que exige uma caracterização do conjunto limite de todo o vector $M_n{}^{(r)}$.

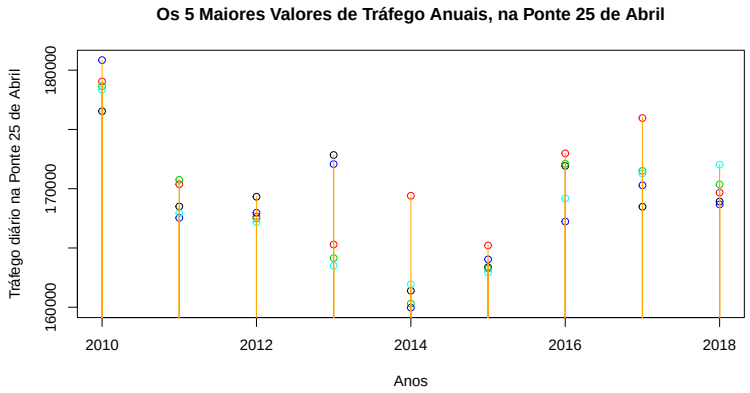


Figura 4.10: Os 5 maiores valores diários de tráfego, por ano, na Ponte 25 de Abril (2010-2018)

Neste caso, os dados serão agrupados em $m = 9$ blocos. Ou seja, no bloco i as maiores observações r_i são gravadas, levando à série $M_i{}^{(r_i)} = \left(z_i{}^{(1)}, \ldots, z_i{}^{(r_i)}\right)$ para $i = 1, \ldots, m$. É usual definir $r_1 = \cdots = r_m = r$ para algum valor de r específico, como já tinha sido referido.

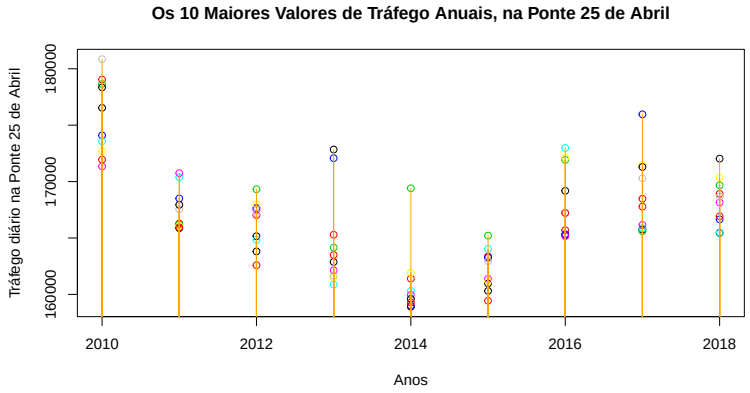


Figura 4.11: Os 10 maiores valores diários de tráfego, por ano, na Ponte 25 de Abril (2010-2018)

A amostra que será utilizada, é composta pelos 3, 5 e 10 maiores valores de tráfego da Ponte 25 de abril para cada um dos anos (2010 até 2018), como já se mencionou. Como tal, a verosimilhança para este modelo é obtida a partir de (3.24) e de (3.25). As estimativas da MV e os erros padrão são dados na tabela 4.6 por inferências baseadas no valor selecionado de r. Não esquecendo que quanto maior o valor de r, mais pequenos os erros padrão, logo correspondem a modelos de maior precisão, mas se a aproximação assintótica é válida para uma escolha de r única, então as estimativas dos parâmetros devem ser estáveis quando o modelo é ajustado com menos estatísticas ordinais.

r	Log-verosimilhança maximizada	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
1	−87.59868	170156.7 (1403.175)	3778.887 (1025.630)	−0.1324375 (0.2395907)
3	−244.9446	172548.763 (1255.148)	4071.888 (526.632)	−0.314111 (0.1484964)
5	−388.9228	172390.0 (1000.367)	3546.707 (356.307)	−0.2503005 (0.09748432)
10	−722.9902	173685.0 (892.955)	3465.030 (468.824)	−0.2356624 (0.0831861)

Tabela 4.6: A log-verosimilhança maximizada, a estimação dos parâmetros e os erros padrão correspondentes, quando considerados os $r = 1, 3, 5$ e 10 maiores valores de tráfego anuais na Ponte 25 de Abril

Nota-se que à medida que se aumenta o número de observações retidas, os erros padrão das estimativas têm tendência para diminuir, exceto num caso para o parâmetro escala que é melhor no caso de $r = 5$ que no caso de $r = 10$. Tendo em conta os valores apresentados na tabela 4.6 o $r = 10$ é o valor que possui os erros padrão das estimativas com menores valores, sem contar com o parâmetro escala.

Em qualquer um dos casos $\hat{\xi} < 0$, logo a distribuição subjacente a estes valores de tráfego diário da Ponte 25 de Abril poderá ser Weibull. Se for esse o caso, a distribuição terá uma cauda leve e com limite superior do suporte finito. Contudo, as estimativas do parâmetro de forma estão muito perto do zero, logo a hipótese da distribuição de Gumbel não deve ser excluída.

r	IC de 95%		
	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
1	[167406.4,172906.9]	[1768.652,5789.123]	[−0.6020352,0.3371602]
3	[170088.7,175008.9]	[3039.689,5104.087]	[−0.60516397, −0.02305811]
5	[170429.3,174350.8]	[2848.346,4245.068]	[−0.44136974, −0.05923121]
10	[171934.8,175435.2]	[2546.135,4383.925]	[−0.39870718, −0.07261767]

Tabela 4.7: Os valores dos IC dos parâmetros estimados pela MV correspondentes, quando considerados os $r = 1, 3, 5$ e 10 maiores valores de tráfego anuais na Ponte 25 de Abril

Observando-se os valores da tabela 4.6 e também da tabela 4.7, verifica-se que existe uma certa estabilidade nas estimativas dos parâmetros de localização e de escala, mesmo que a variabilidade de amostragem seja contabilizada. Isto levanta a dúvida da validade do modelo para valores de $r > 5$.

Já que os parâmetros μ, σ e ξ correspondem exatamente aos parâmetros do modelo

GEV da distribuição de máximos anuais, para se avaliar o ajuste do modelo com mais detalhe conseguem-se derivar as curvas do NR da distribuição dos máximos anuais. São efetuadas do mesmo modo que o modelo GEV, no entanto, neste caso utilizam-se as estimativas de MV e a matriz da variância-covariância do modelo estatístico das r maiores observações. Na figura 4.12 têm-se os gráficos para cada valor de r de 2 a 10. Na mesma figura verifica-se que a concordância entre o modelo e os dados, à medida que o r aumenta, vai diminuindo.

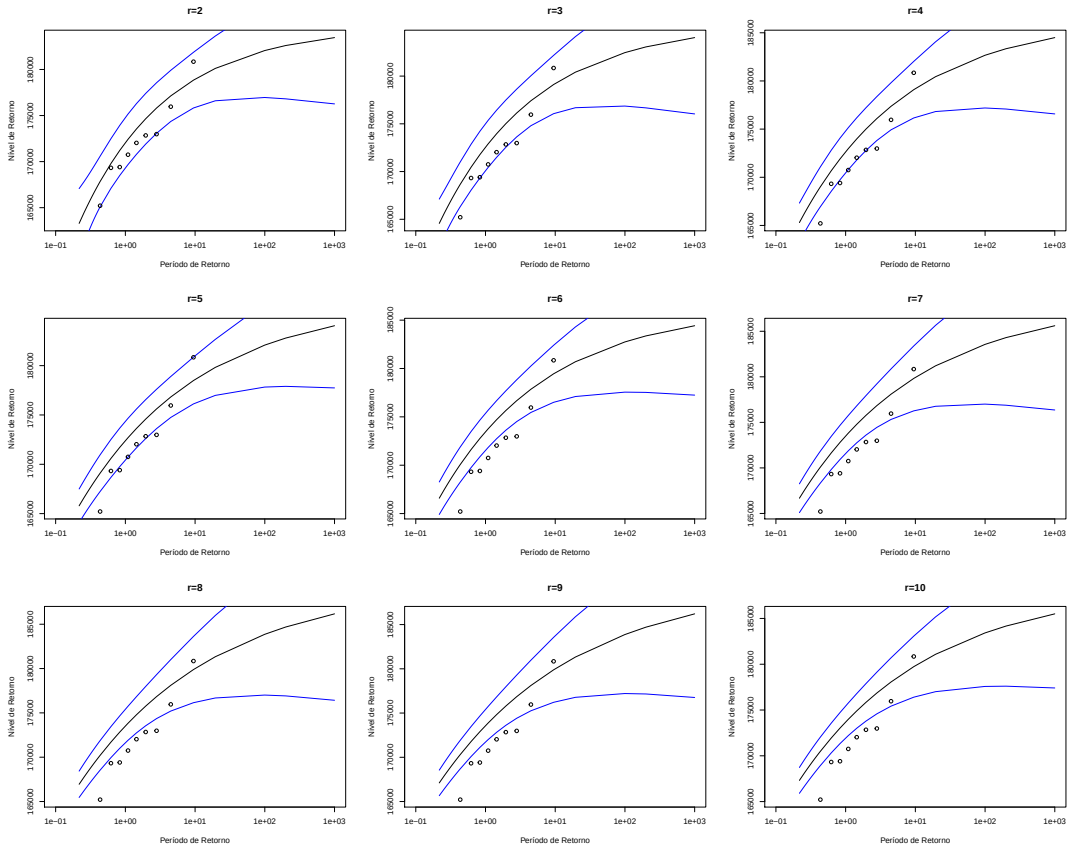


Figura 4.12: Os NR estimados com IC de 95% para a distribuição de máximos anuais baseados no Modelo estatístico das r maiores observações ajustado aos dados do tráfego da Ponte 25 de Abril

Para qualquer escolha de r, a precisão do ajuste pode ser verificada com maior detalhe, para cada $r = 3$, 5 e 10 o conjunto usual de disgnóstico é mostrado nas figuras 4.13, 4.14 e 4.15. Relativamente aos gráficos de NR, estes obtêm-se exatamente do mesmo modo que para o modelo de blocos de máximos, substituindo as estimativas de parâmetros e a matriz de variância-covariância pelas obtidas pela maximização de (3.26). Como se pode verificar as diferenças são bastante acentuadas, nota-se que o melhor ajuste para o máximo anual do tráfego da Ponte 25 de Abril ocorre quando são tidas em conta as 3 maiores observações em cada ano. Consegue-se tirar essa conclusão pela observação do gráfico do NR, em que se verifica que para este caso os valores mostrados encontram-se

maioritariamente dentro das linhas azuis dos IC, no caso em que $r = 5$ os pontos já se encontram um pouco mais afastados e no caso de $r = 10$ é raro o ponto que se encontre dentro dos limites desenhados pelo IC.

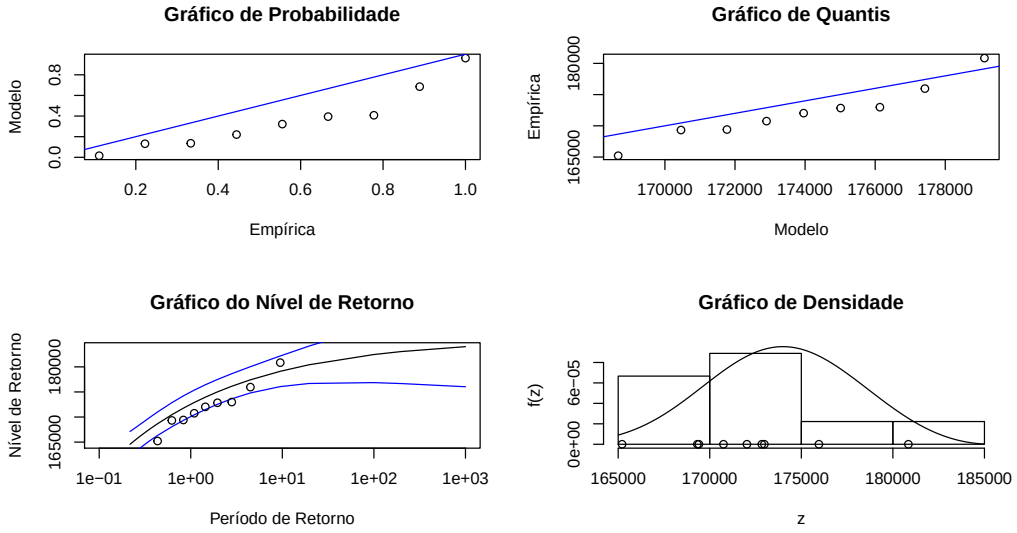


Figura 4.13: Gráficos diagnóstico do Modelo estatístico das r maiores observações com $r = 3$ para os maiores valores anuais de tráfego na Ponte 25 de Abril

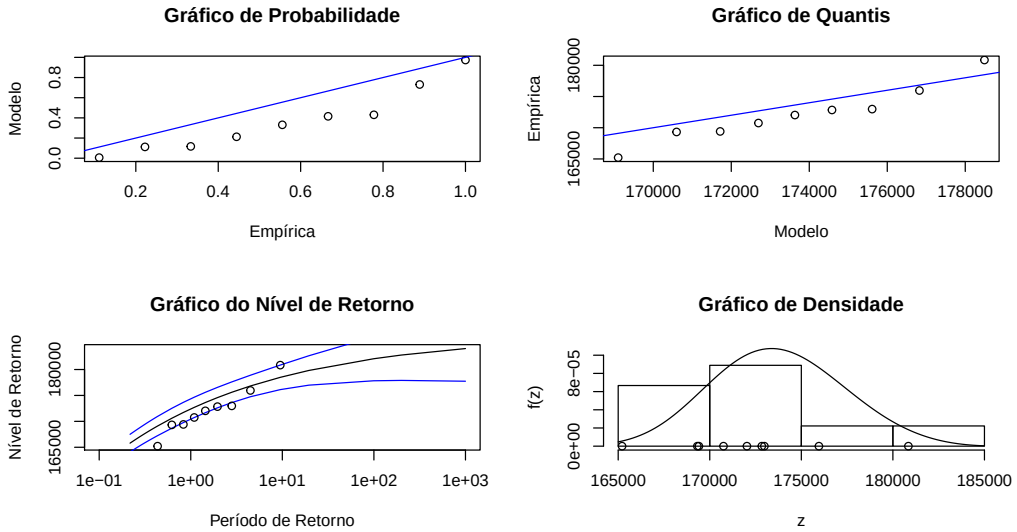


Figura 4.14: Gráficos diagnóstico do Modelo estatístico das r maiores observações com $r = 5$ para os maiores valores anuais de tráfego na Ponte 25 de Abril

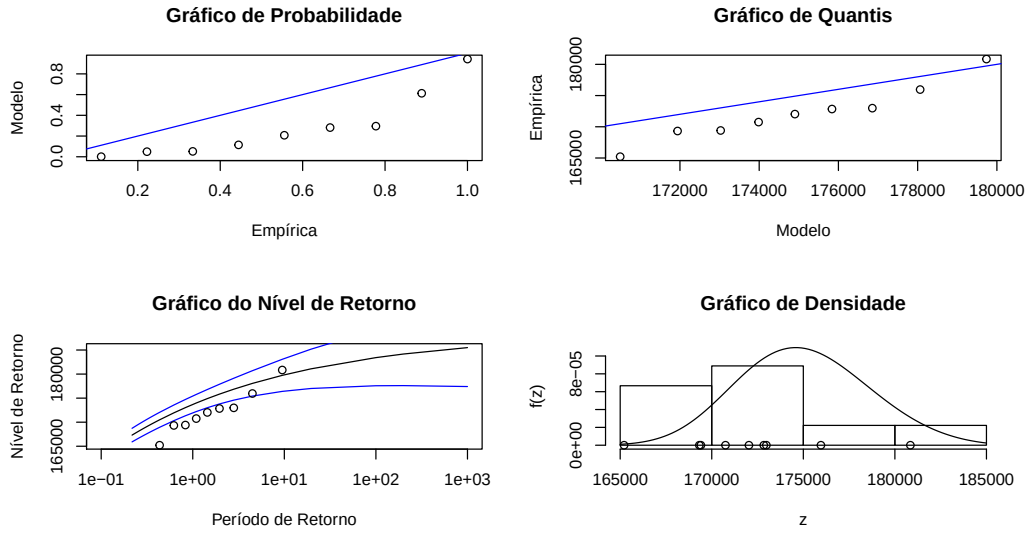


Figura 4.15: Gráficos diagnóstico do Modelo estatístico das r maiores observações com $r = 10$ para os maiores valores anuais de tráfego na Ponte 25 de Abril

Para os dados do tráfego da Ponte 25 de Abril, a preocupação com a falta de ajuste é reforçada pelos gráficos das figuras 4.13, 4.14 e 4.15. As verificações também podem ser feitas sobre a qualidade do ajuste para cada uma das estatísticas do pedido, fazendo gráficos de probabilidade e de quantis. Estes são obtidos comparando a distribuição da estatística de k ordem, (3.23), com os valores dos parâmetros substituídos pelas suas estimativas (com as estimativas empíricas correspondentes).

Para os dados do tráfego da Ponte 25 de Abril, com o modelo ajustado correspondente a $r = 5$, os gráficos de probabilidade e de quantis para cada uma das 4 estatísticas de maiores observações são dadas pela figura I.3 em anexo. Estes gráficos mostram alguma falha na adequação ao modelo.

Na tabela 4.8 encontram-se as estimativas de NR de 5, 10, 50 e 100 anos, para cada um dos valores de r, portanto, para $r = 1$, $r = 3$, $r = 5$ e $r = 10$ e ainda os IC de, aproximadamente, 95% que foram calculados pelo método delta.

		$r=1$	$r=3$	$r=5$	$r=10$
5 anos	$\hat{z}_{0.2}$ **IC de 95%**	175297 [171770,178824]	177419 [174802,180036]	176825 [174750,178900]	178063 [175424,180702]
10 anos	$\hat{z}_{0.1}$ **IC de 95%**	177510 [173121,181900]	179119 [176065,182173]	178492 [176122,180862]	179737 [176401,183072]
50 anos	$\hat{z}_{0.02}$ **IC de 95%**	181671 [173267.9,190074.8]	181706 [176930,186483]	181224 [177610,184838]	182526 [177435,187617]
100 anos	$\hat{z}_{0.01}$ **IC de 95%**	183174 [172367,193982]	182456 [176861,188051]	182080 [177828,186332]	183415 [177575,189256]
$\hat{z}_0$ **IC de 95%**		198690 [104411,292969]	185512 [173339,197685]	186560 [175071,198049]	188388 [174602,202175]

Tabela 4.8: Valores dos NR e dos IC quando $r = 1$, 3, 5 e 10 maiores valores de tráfego anuais na Ponte 25 de Abril

Já que nos casos apresentados $\hat{\xi} < 0$, também é possível fazer inferências sobre o limite superior do suporte da distribuição que é efetivamente o 'período inferior de retorno da observação', isto é, $\hat{z}_0$. Este valor está calculado na última linha da tabela 4.8 e os seus respetivos IC de 95% (aproximadamente). O $\hat{z}_0$ como seria de esperar é o maior valor para $\hat{z}_p$ e os IC de $\hat{z}_0$ são os que possuem maior amplitude, os dois limites destes intervalos são, efetivamente, o menor e o maior valor obtidos.

Depois de se verificarem os valores obtidos para estimativas dos parâmetros, para os erros padrão e observando-se os gráficos diagnóstico (4.13, 4.14 e 4.15) resultantes do ajuste do Modelo estatístico para as $r = 3$, 5 e 10 maiores observações dos valores de tráfego anuais na Ponte 25 de Abril, pode-se concluir que o ajuste efetuado que aparenta ser ligeiramente melhor que os restantes para estes dados é o Modelo estatístico das $r = 5$ maiores observações.

4.4 Modelo GP

Nesta parte, vai-se seguir o que foi descrito na secção 3.4.

4.4.1 Seleção do limiar

O Teorema 11 sugere um modo para a modelagem das maiores observações. Neste caso, os dados são os valores diários do tráfego da Ponte 25 de Abril, representados por uma sucessão de medidas $x_i, \ldots, x_n$. Os eventos extremos serão identificados por um limiar u, para o qual as excedências são $\{x_i : x_i > u\}$ e representam-se por $x_{(1)}, \ldots, x_{(k)}$.

Como mencionado na secção 3.4.3.1 uma forma que ajuda a saber qual o limiar u que se deve selecionar é a visualização do GVRM. Na figura 4.16 está representado o mesmo referente aos dados aqui abordados.

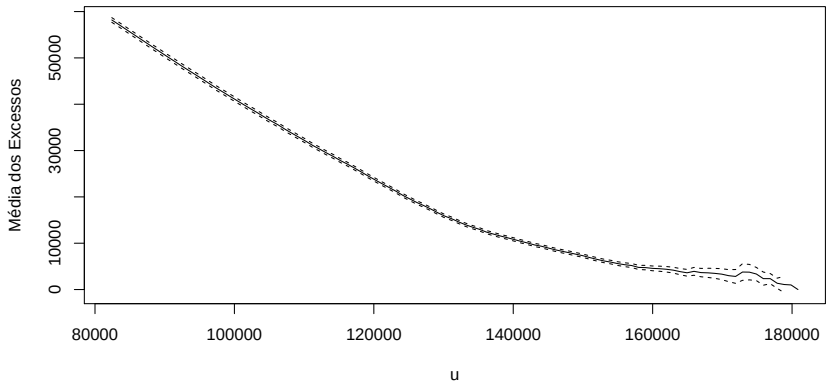


Figura 4.16: GVRM para os dados do tráfego diário da Ponte 25 de Abril

Ao observar o gráfico 4.16 verifica-se que existe uma relação inversa entre a quantidade de valores e a grandeza dos mesmos, isto é, quanto maior é a quantidade existente mais pequenos são os valores de tráfego diário representados e a quantidade vai diminuindo quando o respectivo valor representado vai aumentando. No entanto, este decréscimo nem sempre é igual, ou seja, é mais acentuado até cerca do ponto $u = 135000$, depois varia a inclinação, continuando a diminuir, desta vez, com um ritmo menos acelerado e existe um ponto a partir do qual o declive diminui mais um pouco.

Quanto à seleção de u: a prática standard é adotar como limiar o valor mais baixo possível, que levará, em princípio, ao ajuste de um modelo limiar que irá fornecer uma aproximação razoável.

Para estes dados em concreto foram escolhidos três possíveis valores para o u que serão comparados. O primeiro valor é $u = 165212$, representa o valor mínimo dos máximos anuais. O segundo valor selecionado para o u é 156297, é o valor encontrado através do GVRM onde o comportamento do gráfico mais se altera. Por último, o terceiro valor selecionado para o u é 161734, onde se teve em conta a prática standard e se selecionou o u cujas excedências correspondessem a 5% do valor total da amostra.

De seguida na figura 4.17, assinalam-se os lugares dos valores dos limiares u selecionados em cada um dos três casos referidos, com uma linha vermelha, uma linha azul e uma linha verde.

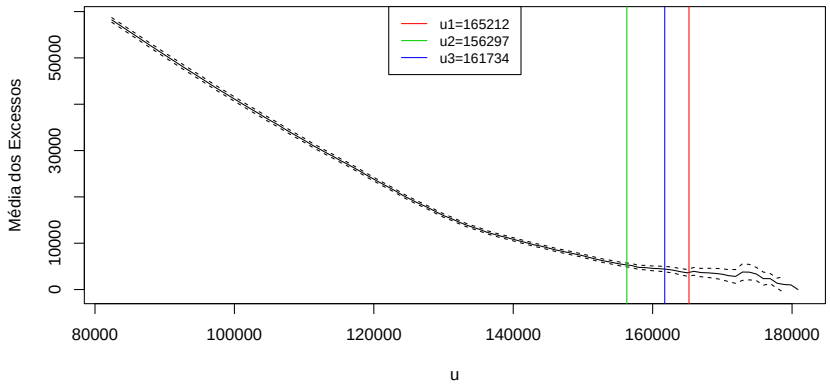


Figura 4.17: GVRM com o lugar dos valores dos limiares representados com cores diferentes para os dados do tráfego diário da Ponte 25 de Abril

4.4.2 Estimação de Parâmetros

Para se calcularem os valores da estimativa de MV, usou-se a função *gpd.fit()* do pacote que se tem utilizado, o cálculo foi efetuado para cada um dos valores de u. Os *outputs* correspondentes foram os três seguintes:

1. Para $u1 = 165212$:

R code 4.3: *Output do gpd.fit() para* $u1 = 165212$

```
$threshold
[1] 165212

$nexc
[1] 84

$conv
[1] 0

$nllh
[1] 773.3564

$mle
[1] 3876.91407301   -0.05639767

$rate
[1] 0.02555522

$se
[1] 662.6056307   0.1304004
```

2. Para $u2 = 156297$:

R code 4.4: *Output do gpd.fit() para* $u2 = 156297$

```
$threshold
[1] 156297

$nexc
[1] 430

$conv
[1] 0

$nllh
[1] 4113.271

$mle
[1] 6354.6384655   -0.1912576

$rate
[1] 0.1308184

$se
[1] 393.43799623   0.03851097
```

3. Para $u3 = 161734$:

R code 4.5: *Output do gpd.fit() para* $u3 = 161734$

```
$threshold
[1] 161734

$nexc
[1] 165

$conv
[1] 0

$nllh
[1] 1548.466

$mle
[1] 5179.9469359   -0.1677628

$rate
[1] 0.05019775
```

```
18
19 $se
20 [1] 538.65428333   0.06930568
```

As Estimativas da MV para os parâmetros e os respetivos IC de aproximadamente 95%, para cada limiar u estão representados na seguinte tabela 4.9.

Limiar u	Valores dos parâmetros $(\hat{\sigma}, \hat{\xi})$	IC de 95%	
		IC para $\hat{\sigma}$	IC para $\hat{\xi}$
165212	(3876.914, −0.0563998)	[2578.207,5157.621]	[−0.31198,0.19919]
156297	(6354.638, −0.19126)	[5583.50,7125.777]	[−0.26674, −0.11578]
161734	(5179.947, −0.16776)	[4124.185,6235.709]	[−0.30360, −0.03192]

Tabela 4.9: Valores estimados dos parâmetros e respetivos IC, para diferentes limiares

Como se pode verificar através da tabela 4.9 a estimativa do parâmetro de forma é sempre menor que zero quando estimado e os respetivos IC contêm maioritariamente valores negativos. Na tabela 4.10 estão os valores das log-verosimilhanças maximizadas e respetivas matrizes de variância-covariância para cada limiar.

Limiar u	Log-verosimilhança maximizada	Matriz variância-covariância
165212	−773.3564	$\begin{bmatrix} 439046.222 & -68.8926 \\ -68.8926 & 0.017004 \end{bmatrix}$
156297	−4113.271	$\begin{bmatrix} 154793.457 & -11.9417 \\ -11.9417 & 0.001483 \end{bmatrix}$
161734	−1548.466	$\begin{bmatrix} 290148.437 & -29.4757 \\ -29.4757 & 0.004803 \end{bmatrix}$

Tabela 4.10: A log-verosimilhança maximizada e a matriz variância-covariância estimadas para os dois parâmetros, para os diferentes limiares

4.4.3 Verificação do modelo

Os gráficos de probabilidade, de quantis, de NR e de densidade são todos úteis para avaliar a qualidade do ajuste do modelo GP.

Apresentam-se a seguir os gráficos diagnóstico para o modelo ajustado GP para cada um dos limiares representados nas figuras 4.18, 4.19 e 4.20. Não esquecendo que os gráficos de probabilidade e quantis devem consistir em pontos que são aproximadamente lineares quando um modelo GP é razoável para modelar as excedências de u. Já a função de densidade do modelo GP ajustado é comparada com um histograma das excedências dos limiares.

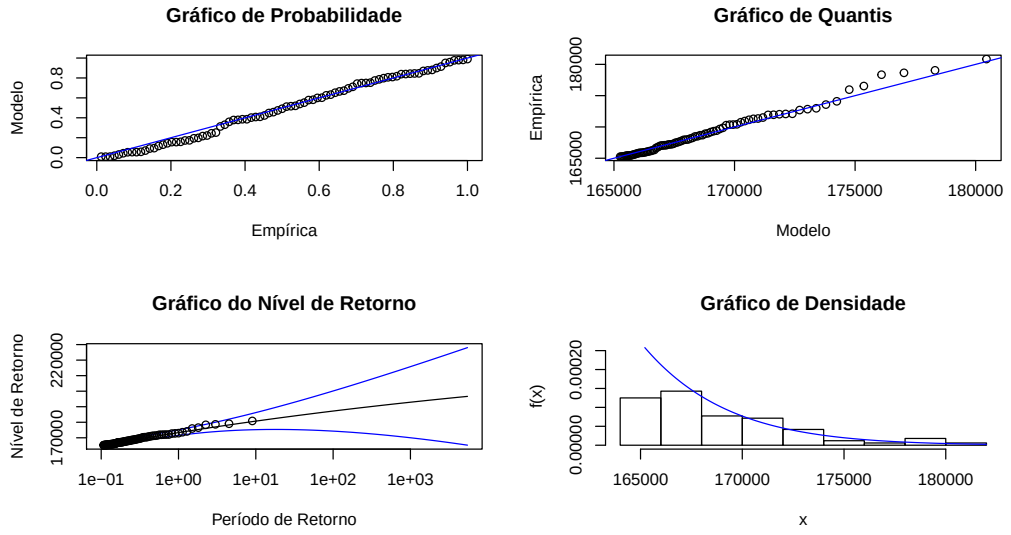


Figura 4.18: Gráficos diagnóstico para o modelo ajustado ao primeiro limiar, $u1 = 165212$

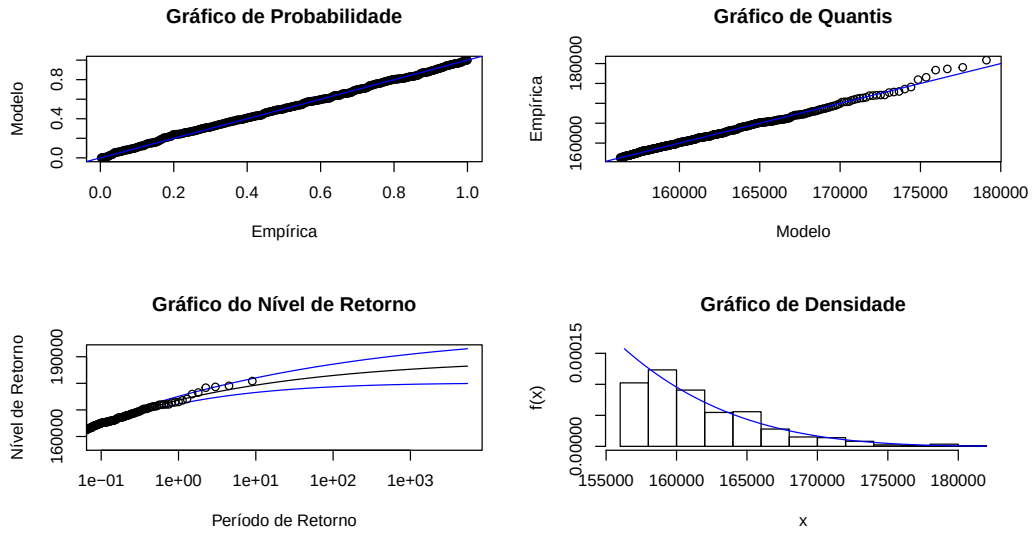


Figura 4.19: Gráficos diagnóstico para o modelo ajustado ao segundo limiar, $u2 = 156297$

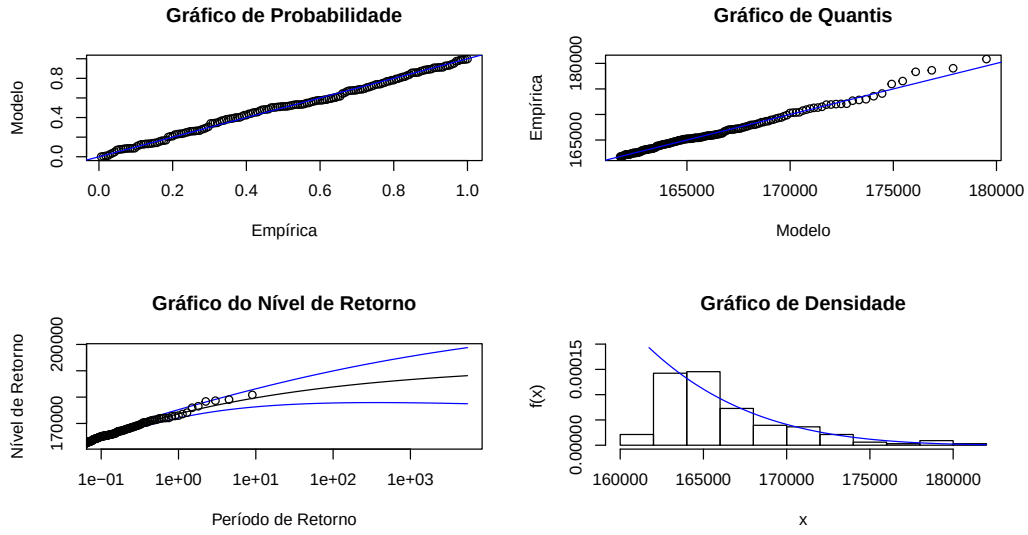


Figura 4.20: Gráficos diagnóstico para o modelo ajustado ao terceiro limiar, $u3 = 161734$

Depois de observados os gráficos diagnóstico pode afirmar-se que, tendo em conta o que já foi mencionado sobre os gráficos de probabilidade e quantil e observando unicamente estes dois gráficos, o segundo limiar é o que possui os valores com um comportamento mais linear. Em relação ao gráfico do NR, o primeiro limiar é o que tem um gráfico com um comportamento mais satisfatório, já que os pontos se encontram todos entre os limites de confiança.

4.4.4 Níveis de retorno

Como já foi mencionado, é conveniente interpretar modelos de valores extremos em termos de quantis ou NR, em vez de valores de parâmetros individuais. Para isso calcularam-se as excedências para cada limiar u no conjunto completo das 3287 observações, também se efetuou o cálculo da estimativa da MV da probabilidade de excedências; da respetiva variância e da matriz variância-covariância para $(\hat{\zeta}_u, \hat{\sigma}, \hat{\xi})$. Os resultados obtidos estão representados na tabela 4.11.

Limiar u	Excedências ao limiar	$\hat{\zeta}_u$	$Var(\hat{\zeta}_u)$	Matriz variância-covariância
165212	84	2.56%	7.58×10^{-6}	$\begin{bmatrix} 7.56 \times 10^{-6} & 0 & 0 \\ 0 & 439046.222 & -68.8926 \\ 0 & -68.8926 & 0.017004 \end{bmatrix}$
156297	430	13.08%	3.46×10^{-5}	$\begin{bmatrix} 3.46 \times 10^{-5} & 0 & 0 \\ 0 & 154793.457 & -11.9417 \\ 0 & -11.9417 & 0.001483 \end{bmatrix}$
161734	165	5.02%	1.45×10^{-5}	$\begin{bmatrix} 1.45 \times 10^{-5} & 0 & 0 \\ 0 & 290148.437 & -29.4757 \\ 0 & -29.4757 & 0.004803 \end{bmatrix}$

Tabela 4.11: Valores: das excedências ao limiar; da probabilidade de excedência; variância; matriz variância-covariância para os três parâmetros com diferentes limiares

Como é mais conveniente mostrar os NR numa escala anual, de tal modo que o NR do ano N é o nível excedido em média uma vez a cada N anos, foram calculados os NR para 5,10, 50 e 100 anos, para cada um dos limiares. Sendo que, por exemplo, o NR de 5 anos corresponde ao NR da observação m com $m = 356 \times 5 = 1825$.

Ao se substituir na (3.40) obtêm-se os valores de $\hat{x}_m$, ou seja, do NR da observação m, e ao se substituir na (3.42) obtém-se pelo método delta o valor da $Var(\hat{x}_m)$, deste modo, será possível calcular também um IC de, aproximadamente, 95% para x_m. Os valores obtidos, para cada um dos limiares, estão representados nas três tabelas seguintes:4.12;4.13;4.14.

Limiar $u = 165212$				
N	$m = N \times 365$	$\hat{x}_m$	$Var(\hat{x}_m)$	IC 95%
5	1825	178605	4007351	[174681,182529]
10	3650	180727	7857282	[175233,186221]
50	18250	185346	28000777	[174974,195717]
100	36500	187209	43434289	[174292,200127]

Tabela 4.12: Valores obtidos para diferentes anos de NR para o primeiro limiar

Limiar $u = 156297$				
N	$m = N \times 365$	$\hat{x}_m$	$Var(\hat{x}_m)$	IC 95%
5	1825	177863	1377036	[175563,180163]
10	3650	179311	1954302	[176571,182051]
50	18250	182016	3771646	[178210,185823]
100	36500	182948	4725744	[178688,187209]

Tabela 4.13: Valores obtidos para diferentes anos de NR para o segundo limiar

Limiar $u = 161374$				
N	$m = N \times 365$	$\hat{x}_m$	$Var(\hat{x}_m)$	IC 95%
5	1825	178140	1939890	[175410,180870]
10	3650	179729	3046671	[176307,183150]
50	18250	182777	7188168	[177522,188032]
100	36500	183856	9646395	[177769,189944]

Tabela 4.14: Valores obtidos para diferentes anos de NR para o terceiro limiar

4.4.5 Escolha do limiar revista

Como explicado na secção 3.4.3.4 uma técnica complementar é ajustar a distribuição GP numa gama de limiares e procurar a estabilidade das estimativas dos parâmetros. Os gráficos de $\hat{\sigma}^*$ e $\hat{\xi}$ contra u são os dois na figura 4.21.

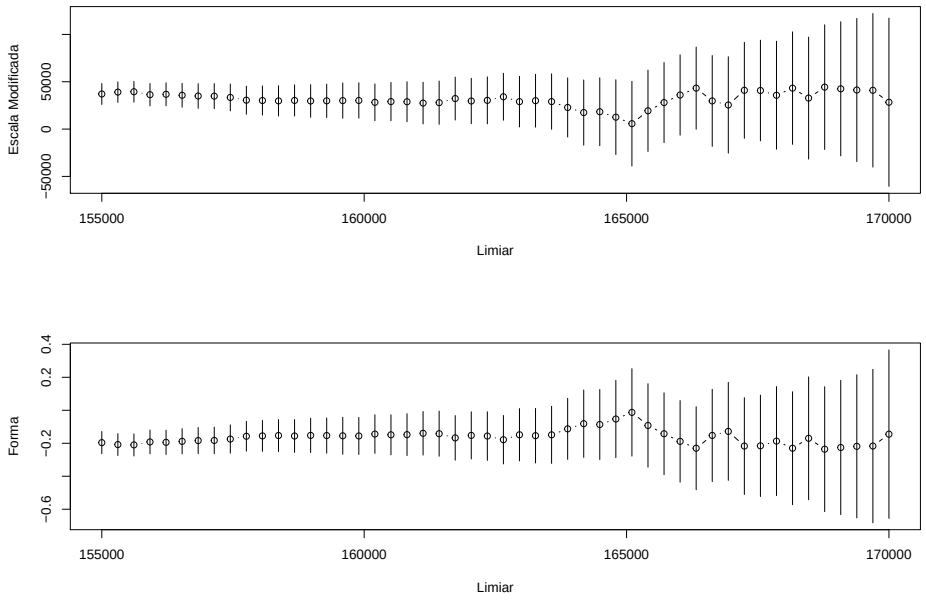
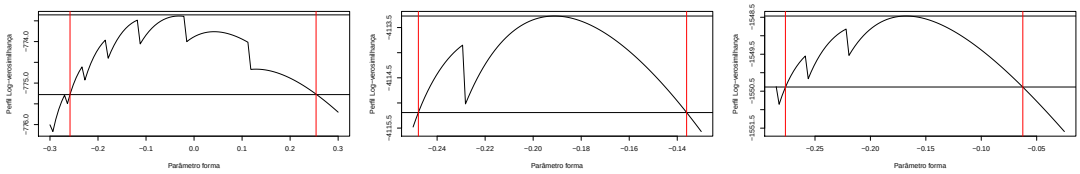


Figura 4.21: Estimação de parâmetros para 50 limiares diferentes para os dados diários do tráfego da Ponte 25 de Abril

Só foi possível fazer o gráfico para um intervalo mais pequeno de valores, como se vê, calculou-se de 155000 até 170000. Por isso, selecionou-se o intervalo de valores que, segundo o observado no GVRM, seriam os de maior relevância. E, tal como observado no gráfico 4.16, o padrão de mudança para limiares muito altos também é patente nesta representação 4.21, mas, neste último, as perturbações parecem pequenas em relação aos

erros de amostragem. Aparentemente, tendo em conta a figura 4.21, o primeiro valor do u será o mais razoável.

A melhor precisão é obtida utilizando os IC do perfil da log-verisimilhança. As figuras seguintes mostram o perfil da log-verosimilhança para ξ, para os diferentes limiares.

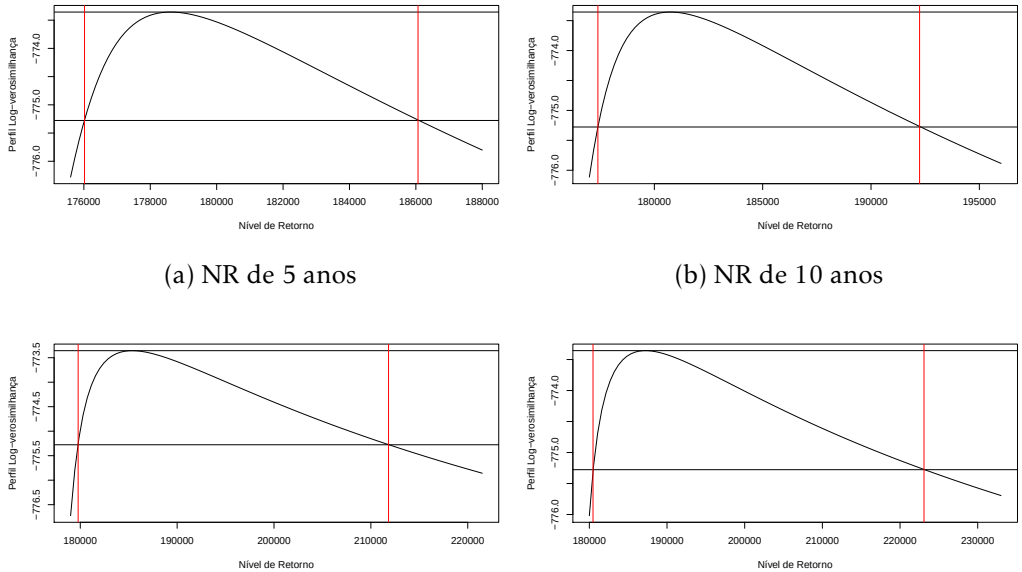


(a) $u1 = 165212$ (b) $u2 = 156297$ (c) $u3 = 161734$

Figura 4.22: Gráficos do perfil da log-verosimilhança para ξ, no modelo de excedências do limiar, aplicados nos dados do tráfego da Ponte 25 de Abril

Um IC de 95%, aproximadamente, para ξ, é obtido a partir dos gráficos como $[-0.2582, 0.2540]$ para o primeiro limiar; $[-0.2478, -0.1361]$; para o segundo limiar e $[-0.2767, -0.0625]$ para o terceiro limiar.

Já o perfil da log-verosimilhança para os NR de diferentes anos são representados nos gráficos seguintes, para os três limiares selecionados.

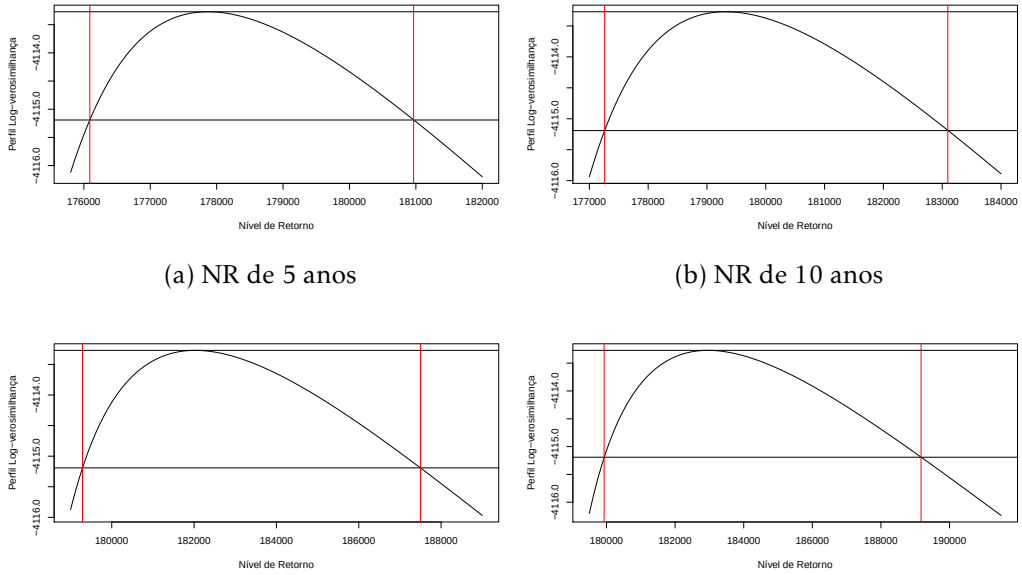


(a) NR de 5 anos (b) NR de 10 anos

(c) NR de 50 anos (d) NR de 100 anos

Figura 4.23: Gráficos dos NR para anos diferentes, para o primeiro limiar, $u1 = 165212$

O IC de 95%, aproximadamente, para o NR de 5 anos é obtido a partir do perfil da log-verosimilhança como $[176020, 186065]$; para 10 anos é $[177395, 192242]$; para 50 anos é $[179780, 211825]$; para 100 anos é $[180475, 223085]$.



(a) NR de 5 anos

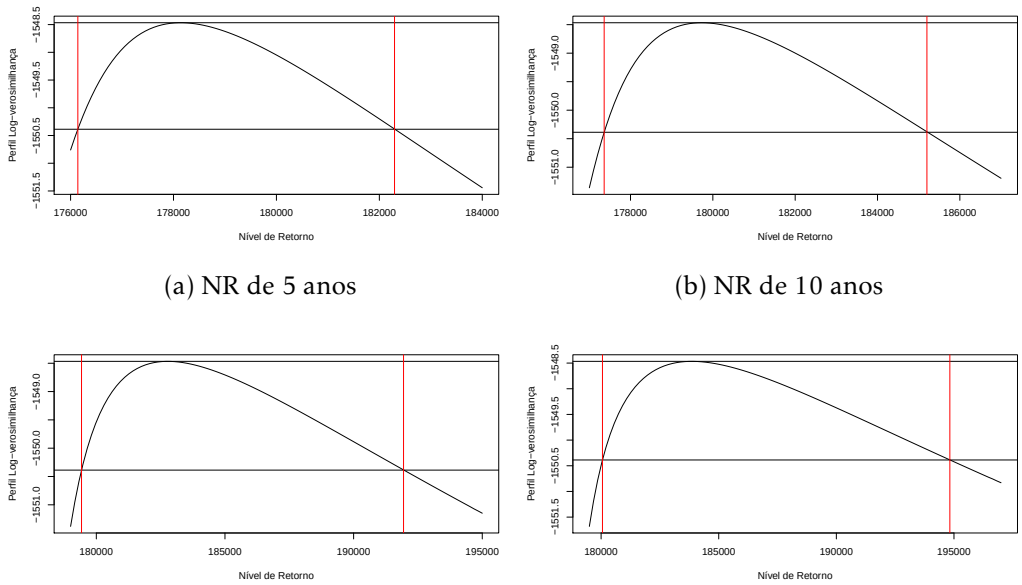
(b) NR de 10 anos

(c) NR de 50 anos

(d) NR de 100 anos

Figura 4.24: Gráficos dos NR para anos diferentes, para o segundo limiar, $u2 = 156297$

O IC de 95%, aproximadamente, para o NR de 5 anos é obtido a partir do perfil da log-verosimilhança como [176090, 180966]; para 10 anos é [177260, 183095]; para 50 anos é [179289, 187500]; para 100 anos é [179930, 189170].



(a) NR de 5 anos

(b) NR de 10 anos

(c) NR de 50 anos

(d) NR de 100 anos

Figura 4.25: Gráficos dos NR para anos diferentes, para o terceiro limiar, $u3 = 161734$

O IC de 95%, aproximadamente, para o NR de 5 anos é obtido a partir do perfil da

log-verosimilhança como [166452, 189828]; para 10 anos é [167522, 191935]; para 50 anos é [170027, 195526]; para 100 anos é [171082, 196630].

CAPÍTULO 5

Conclusões e problemas por analisar

A Teoria de Valores Extremos é uma vertente da Estatística por explorar e consegue estudar o que outras áreas ignoram, que são os valores extremos, tantas vezes considerados como "outliers" que "estragam as previsões" e, no entanto, são normalmente os valores que têm mais impacto tanto a nível ambiental (como numa catástrofe natural), como a nível financeiro quando existem "crashs" ou "booms" no mercado da bolsa.

Este ramo da Estatística não vai impedir que estes casos ocorram contudo pode ajudar a prever e compreender estes acontecimentos e, deste modo, permitir minorar ou evitar que as consequências sejam tão catastróficas.

Nesta dissertação foi elaborada uma análise da aplicação da Teoria dos Valores Extremos ao tráfego da Ponte 25 de Abril, um dos locais com maior fluxo de tráfego diário recorrente em todo o país. Estimaram-se os parâmetros do modelo em questão, e fizeram-se inferências sobre os níveis de retorno para um determinado número de anos, sobre os períodos de retorno, etc., que são fulcrais para a previsão de fluxos de grande tráfego.

Relembrado, a amostra original disponibilizada é constituída por: registos diários do tráfego da Ponte 25 de Abril desde 1 de janeiro de 2010 até 31 de dezembro de 2018; registos do tráfego médio diário mensal desde 2006; registos do tráfego médio diário anual desde 1966, de que foram efetuadas sub-amostras para se aplicarem os Modelos da Teoria dos Valores Extremos, como por exemplo, os valores máximos anuais, com o objetivo de serem aplicados aos máximos agrupados em blocos e à distribuição Gumbel.

Os métodos aplicados aos valores máximos anuais mostraram que o parâmetro de forma, ξ, apresentou valores menores que zero, o que significa que a distribuição subjacente aos valores do tráfego anual máximo poderá ser a distribuição Weibull. No entanto, a distribuição Gumbel não poderá ficar de lado visto que houve IC que incluíam o zero. Como se pode observar na tabela 4.2 o IC do parâmetro forma é maioritariamente negativo, pelo que a distribuição subjacente deverá ser Weibull, mas deve também incluir a análise da distribuição Gumbel pelo facto do IC inclui o $\xi = 0$.

No Modelo estatístico das r maiores observações fizeram-se três sub-amostras, com as

três, as cinco e as dez maiores observações de cada ano (desde 2010 a 2018), tendo em consideração, não só as estimativas dos parâmetros e dos erros padrão das estimativas para cada um dos valores de r, como também, a observação dos gráficos diagnóstico, 4.13, 4.14 e 4.15. A qualidade do ajuste para os máximos anuais do fluxo de tráfego na Ponte 25 de Abril, parece ser ligeiramente melhor quando são retidas as 5 maiores observações em cada ano.

No método do modelo GP selecionaram-se os valores acima de três limiares diferentes, tendo presente os gráficos diagnóstico, 4.18, 4.19 e 4.20. Ao visualizar, de modo mais detalhado, os gráficos de probabilidade e de quantil, o melhor valor para u é 156297, ou seja, o segundo limiar. Este também é o valor cujos erros padrão das estimativas são menores.

Nesta tese focou-se o estudo no tráfego da Ponte 25 de Abril, mas seria relevante fazer o mesmo estudo noutras Pontes, principalmente, na Ponte Vasco da Gama, fazendo uma relação entre as duas através de um Modelo para Extremos Bivariados.

Neste estudo utilizaram-se os dados referentes ao tráfego diário, únicos disponibilizados. No entanto, se for possível disponibilizar os dados horários, poder-se-iam fazer as previsões horárias de maior fluxo de tráfego e, com essa informação disponível, tomar decisões quanto às deslocações e utilização da ponte. Quanto às empresas que são responsáveis por estas infraestruturas, poderiam eleger os melhores horários para possíveis manutenções necessárias ou serem tidas em conta para outras ações.

Nesta tese não foi abordado o impacto dos ciclos de carga dos veículos na estrutura da Ponte 25 de Abril. Poderá ter interesse na área da Engenharia das Estruturas e como informação para a Lusoponte. Existem já estudos feitos nesta vertente, como é o caso do artigo Yang, Zhang e Ren (2018).

Em relação à análise financeira da Ponte 25 de Abril, foi efetuado um estudo com base nas receitas cobradas e no valor unitário pago nas Portagens por cada uma das Classes. Verificou-se que tem havido um aumento dos preços unitários, por Classe, ao longo dos anos, pelo menos, desde 1996, bem como das receitas recolhidas. Ainda se verificou que as receitas da Lusoponte são maioritariamente provenientes da Ponte 25 de Abril. Todavia, não foram exploradas as receitas a nível líquido, nem a percentagem que efetivamente é lucro para a Lusoponte, já que ao ser uma entidade "Público-Privada" tem um modo de funcionar distinto em relação ao fim das receitas.

Referências Bibliográficas

Almeida, I. (2018, novembro 7). Ponte 25 de abril não está em risco mas precisa de obras. Obtido de http://www.lisbonne-idee.pt/p5383-ponte-abril-nao-esta-risco-mas-precisa-obras.html

Beirlant, J., Goegebeur, Y., Segers, J. & Teugels, J. L. (2006). *Statistics of extremes: Theory and applications*. John Wiley e Sons Ltd.

Bureau, U. S. C. (2017, janeiro 18). *X-13arima-seats reference manual accessible html output version*.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer Series in Statistics.

Costa, E. (2018, novembro 7). Ponte 25 de abril e tejo. Obtido de https://emanueljccosta.files.wordpress.com/2014/10/ponte-25-de-abril-e-tejo-236.jpg

Ferreira, P. G. C. & Mattos, D. M. (2016). *Usando o r para ensinar ajuste sazonal*. Instituto Brasileiro de Economia (FGV|IBRE).

Fisher, R. A. & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. Em *Mathematical proceedings of the cambridge philosophical society* (Vol. 24, *2*, pp. 180–190). Cambridge University Press.

Garcia, A., Pignatelli, C., Salina, A. & Santos, G. (2000). *Auditoria à aplicação do modelo contratual e aos acordos de reposição do equilíbrio financeiro*. Tribunal de Contas Sector Público Empresarial – DA IX.

GITHUB. (2009, janeiro 13). Obtido de https://github.com/cran/ismev/blob/master/R/gev.R

Gnedenko, B. (1943). Sur ia distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics*, 423–453.

Heffernan, J. E. & Stephenson, A. G. (2018, maio 8). *Ismev: An introduction to statistical modeling of extreme values*.

INE. (2018, novembro 7). Obtido de www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0008937&contexto=bd&selTab=tab2

Infraestruturas de Portugal, S. (2017). *Relatório síntese de execução orçamentar 4º trimestre 2017*.

Infraestruturas de Portugal, S. (2018a, novembro 7). 50 anos da ponte 25 de abril - linha do tempo. Obtido de www.infraestruturasdeportugal.pt/50-anos-da-ponte-25-de-abril/linha-do-tempo

Infraestruturas de Portugal, S. (2018b). *Relatório e contas consolidado 2018 primeiro semestre*.

Jornal de Negócios. (2012, julho 30). Obtido de www.jornaldenegocios.pt/economia/detalhe/mecircs_de_agosto_volta_a_ser_pago_na_ponte_25_de_abril

Leadbetter, M. R., Lindgren, G. & Rootzen, H. (1983). *Extremes and related properties of random sequences and processes*. Springer Verlag.

Lima, F. (2018). *Estatísticas dos transportes e comunicações 2017*.

Lusoponte, C. (2019a, fevereiro 26). Informacoes gerais - ponte 25 de abril. Obtido de www.lusoponte.pt/25-de-abril/informacoes-gerais

Lusoponte, C. (2019b, março 15). Informacoes gerais - ponte vasco da gama. Obtido de www.lusoponte.pt/vasco-da-gama/informacoes-gerais

Maravall, A. (2005). An application of the tramo-seats automatic procedure; direct versus indirect adjustment. *Computational Statistics & Data Analysis*, *50*(9), 2167–2190.

Mises, R. v. (1936). La distribution de la plus grande de n valeurs. *Rev. Math. Union Interbalcanique*, *1*, 141–160.

Penalva, H., Neves, M. & Nunes, S. (2013). Topics in data analysis using r in extreme value theory. *Metodoloski zvezki*, (1).

Público. (2006, agosto 6). Obtido de www.publico.pt/2006/08/06/local/noticia/ponte-25-de-abril-primeira-travessia-do-tejo-em-lisboa-completa-hoje-40-anos-1266400

Rosário, P. A. G. (2013). *Análise de valores extremos para níveis pluviométricos em barcelos* (tese de mestrado, Universidade de Lisboa - Faculdade de Ciências).

Sax, C. & Eddelbuettel, D. (2018, dezembro 20). *Seasonal: R interface to x-13-arima-seats*.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, *72*(1), 67–90.

StackExchange. (2018, fevereiro 26). Obtido de https://stats.stackexchange.com/questions/148573/the-results-and-specifics-from-the-qs-function-in-r

Trainlogistic. (2018, novembro 6). Ponte 25 de abril. Obtido de http://www.trainlogistic.com/pt/Estrutura/ObrasArte/Eixo-NS/pt_25abril.htm

Wikipedia. (2018, fevereiro 27). Autorregressive integrated moving average. Obtido de https://en.wikipedia.org/wiki/Autorregressive_integrated_moving_average

Yang, X., Zhang, J. & Ren, W.-X. (2018). Threshold selection for extreme value estimation of vehicle load effect on bridges. *International journal of distributed sensor networks*, *14*(2), 1–12.

ANEXO I

Anexo

I.1 Ajuste sazonal, resultados detalhados

I.1.1 Estatística QS

Código do R para o cálculo da estatística QS:

```
require(seasonal)
m <- seas(x=trafego)
require(polynom)
x <- trafego
S <- frequency(x)
S
[1] 12
y<-udg(m, "x13mdl")
y
x13mdl
(0 1 1)(0 1 1)

ndif <- max(1, min(2, 2))
dx <- filter(x, polynomial(c(1,-1))^ndif, sides=1)
dx <- window(dx, start=time(x)[ndif+1])
R <- acf(dx, lag.max=S*2, plot=FALSE)$acf[-1,,1][c(S, 2*S)]
if (R[1] <= 0)
+   R[1] <- 0
if (R[2] <= 0)
+   R[2] <- 0
R
[1] 0.7092818 0.6373856
n <- length(dx)
n
[1] 106
QS <- n*(n+2)*(R[1]^2/(n-S) + R[2]^2/(n-2*S))
pvalue <- pchisq(q=QS, df=2, lower.tail=FALSE)
round(c(QS=QS, p.value=pvalue), 4)
```

```
29 QS  p.value
30 117.9867   0.0000
31 qs(m)["qsori",]
32 qs    p-val
33 117.9867   0.0000
```

I.1.2 Previsões do tráfego na Ponte 25 de Abril com o ajuste sazonal

Data	Previsão	Limite Inferior IC	Limite superior IC
jan/19	4101281	3918143	4284419
fev/19	3854536	3649774	4059299
mar/19	4314374	4100248	4528500
abr/19	4281630	4067490	4495770
mai/19	4625369	4402324	4848413
jun/19	4571697	4338591	4804803
jul/19	4987469	4743992	5230946
ago/19	4837331	4584004	5090658
set/19	4577284	4313499	4841070
out/19	4464936	4190220	4739652
nov/19	4198770	3914404	4483137
dez/19	4312279	4017568	4606989
jan/20	4211359	3905353	4517364
fev/20	3940587	3622264	3622264
mar/20	4455309	4124362	4786256
abr/20	4395985	4053811	4738158
mai/20	4739824	4386410	5093238
jun/20	4696185	4330561	5061808
jul/20	5110072	4732070	5488074
ago/20	4960036	4570310	5349761
set/20	4704063	4301873	5106253
out/20	4589830	4175558	5004102
nov/20	4329724	3903079	4756369
dez/20	4447306	4008123	4886489

Figura I.1: Primeira previsão do tráfego na Ponte 25 de Abril com ajuste sazonal, valores correspondentes ao gráfico representado na figura 2.13

Ano/mês	JAN	FEV	MAR	ABR	MAI	JUN	JUL	AGO	SET	OUT	NOV	DEZ
2019	4388701	4395892	4403730	4412215	4421347	4431125	4440763	4450023	4459457	4469065	4478847	4488803
2020	4498933	4509237	4519715	4530366	4541192	4552192	4563366	4574714	4600244	4663029	4766106	4898965

Figura I.2: Segunda previsão do tráfego na Ponte 25 de Abril com ajuste sazonal, valores correspondentes ao gráfico representado na figura 2.14

I.2 Análise das portagens e receitas da Ponte 25 de Abril, valores detalhados

Data de início de aplicação	Classe 1	Classe 2	Classe 3	Classe 4
01/01/1996	0,75 €	1,85 €	2,74 €	3,59 €
01/01/2002	1,00 €	2,55 €	3,75 €	4,85 €
01/01/2003	1,05 €	2,65 €	3,90 €	5,05 €
01/01/2004	1,10 €	2,75 €	4,05 €	5,25 €
01/01/2005	1,15 €	2,85 €	4,15 €	5,40 €
01/01/2006	1,20 €	2,95 €	4,30 €	5,60 €
01/01/2007	1,25 €	3,05 €	4,45 €	5,80 €
01/01/2008	1,30 €	3,15 €	4,55 €	5,95 €
01/01/2009	1,35 €	3,25 €	4,70 €	6,15 €
01/07/2010	1,40 €	3,30 €	4,75 €	6,20 €
01/01/2011	1,45 €	3,40 €	4,85 €	6,35 €
01/01/2012	1,55 €	3,55 €	5,05 €	6,60 €
01/01/2013	1,60 €	3,70 €	5,20 €	6,80 €
01/01/2014	1,65 €	3,75 €	5,25 €	6,85 €
01/01/2016	1,70 €	3,80 €	5,30 €	6,95 €
01/01/2017	1,75 €	3,85 €	5,35 €	7,00 €
01/01/2018	1,80 €	3,95 €	5,45 €	7,10 €
01/01/2019	1,85 €	4,05 €	5,55 €	7,20 €

Tabela I.1: Valor unitário das Portagens da Ponte 25 de Abril, das quatro Classes, de 1996 a 2019

Intervalo de tempo	Classe 1	Classe 2	Classe 3	Classe 4	Média
01-01-1996 a 31-12-2001	Aumento				
01-01-2002 a 31-12-2002	0,25 €	0,70 €	1,01 €	1,26 €	0,81 €
01-01-2003 a 31-12-2003	0,05 €	0,10 €	0,15 €	0,20 €	0,13 €
01-01-2004 a 31-12-2004	0,05 €	0,10 €	0,15 €	0,20 €	0,13 €
01-01-2005 a 31-12-2005	0,05 €	0,10 €	0,10 €	0,15 €	0,10 €
01-01-2006 a 31-12-2006	0,05 €	0,10 €	0,15 €	0,20 €	0,13 €
01-01-2007 a 31-12-2007	0,05 €	0,10 €	0,15 €	0,20 €	0,13 €
01-01-2008 a 31-12-2008	0,05 €	0,10 €	0,10 €	0,15 €	0,10 €
01-01-2009 a 30-06-2010	0,05 €	0,10 €	0,15 €	0,20 €	0,13 €
01-07-2010 a 31-12-2010	0,05 €	0,05 €	0,05 €	0,05 €	0,05 €
01-01-2011 a 31-12-2011	0,05 €	0,10 €	0,10 €	0,15 €	0,10 €
01-01-2012 a 31-12-2012	0,10 €	0,15 €	0,20 €	0,25 €	0,18 €
01-01-2013 a 31-12-2013	0,05 €	0,15 €	0,15 €	0,20 €	0,14 €
01-01-2014 a 31-12-2015	0,05 €	0,05 €	0,05 €	0,05 €	0,05 €
01-01-2016 a 31-12-2016	0,05 €	0,05 €	0,05 €	0,10 €	0,06 €
01-01-2017 a 31-12-2017	0,05 €	0,05 €	0,05 €	0,05 €	0,05 €
01-01-2018 a 31-12-2018	0,05 €	0,10 €	0,10 €	0,10 €	0,09 €
01-01-2019 a 31-12-2019	0,05 €	0,10 €	0,10 €	0,10 €	0,09 €

Tabela I.2: Diferença entre os valores unitários das Portagens da Ponte 25 de Abril, das quatro Classes, de 1996 a 2019

Mês | Ano	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Janeiro	1827	1794	1919	x	x	2537	2662	2803	2903	2971	2914	2914	2900	2978	3076	2921	2960	3022	3119	3308
Fevereiro	1724	1720	1896	x	x	2393	2566	2621	2683	2728	2799	2799	2651	2840	2917	2749	2756	2837	3003	3095
Março	1949	1929	2043	x	x	2656	2768	2873	2983	3132	3053	3175	3065	2481	3165	3000	3162	3252	3332	3469
Abril	1812	1884	1874	x	x	2577	2788	2909	2983	3069	3009	3086	3077	3135	2985	3087	3164	3224	3333	3618
Maio	1922	2019	2113	x	x	2844	2920	3061	3199	3226	3094	3271	3241	3260	3259	3276	3421	3488	3509	3699
Junho	1920	2052	2149	x	x	2796	2940	3095	3059	3233	3147	3243	3188	3320	3278	3378	3357	3486	3658	3828
Julho	2071	2245	2292	x	x	3001	3142	3336	3391	3557	3453	3583	3644	3539	3648	3669	3753	3841	4035	4116
Agosto	-	-	-	x	x	-	-	-	-	-	-	-	-	3382	3510	3664	3691	3731	3996	4069
Setembro	1759	1970	2078	x	x	2752	2946	2997	3089	3237	3098	3235	3266	3225	3249	3344	3331	3450	3686	3778
Outubro	1803	1965	2043	x	x	2675	2794	2882	2972	3182	3067	3143	3074	3124	3030	3100	3288	3206	3454	3691
Novembro	1807	1928	1935	x	x	2545	2691	2773	2837	3008	2890	2936	2931	2841	2820	3035	3013	3123	3258	3413
Dezembro	1812	1937	1901	x	x	2603	2767	2865	2922	3067	2946	3002	3018	3000	2988	3117	3223	3259	3401	3485

Tabela I.3: Receitas em milhares de euros da Ponte 25 de Abril de 1998 a 2010

Na tabela I.3 o valores representados a vermelho não se encontram disponíveis no INE, "INE" (2018), e a cor-de-laranja estão marcados os meses de agosto, quando as portagens não eram cobradas, logo estes valores são igual a zero.

Mês | Ano	2011	2012	2013	2014	2015	2016	2017
Janeiro	2978	3076	2921	2960	3022	3119	3308
Fevereiro	2840	2917	2749	2756	2837	3003	3095
Março	2481	3165	3000	3162	3252	3332	3469
Abril	3135	2985	3087	3164	3224	3333	3618
Maio	3260	3259	3276	3421	3488	3509	3699
Junho	3320	3278	3378	3357	3486	3658	3828
Julho	3539	3648	3669	3753	3841	4035	4116
Agosto	3382	3510	3664	3691	3731	3996	4069
Setembro	3225	3249	3344	3331	3450	3686	3778
Outubro	3124	3030	3100	3288	3206	3454	3691
Novembro	2841	2820	3035	3013	3123	3258	3413
Dezembro	3000	2988	3117	3223	3259	3401	3485

Tabela I.4: Receitas em milhares de euros da Ponte 25 de Abril de 2011 a 2017

Ano	Receitas com inflação	Taxa de Inflação (%)	Receitas sem inflação
2003	29379	4,4	29379
2004	30984	3,5	29936
2005	32215	5,8	29419
2006	33021	5,5	28583
2007	34410	1,6	29317
2008	33470	1,5	28094
2009	34387	-3,6	29942
2010	34055	4,6	28349
2011	37125	8,9	28379
2012	37925	3,3	28064
2013	38340	-2,3	29039
2014	39119	-1,2	29989
2015	39919	-1,0	30911
2016	41784	-0,6	32551
2017	43569	3,1	32921

Tabela I.5: Valores das receitas cobradas com e sem inflação a preços constantes de 2003 e a respetiva taxa em cada ano de 2003 a 2017

Os Valores das Receitas sem inflação foram ajustados aos preços de 2003. Para se

efetuar o cálculo das Receitas sem inflação foi dividido o valor das Receitas do ano N por 1 mais a taxa de inflação que aparece no ano $N+1$. Já que a taxa de inflação que aparece no ano N se refere ao ano $N-1$.

Ano	Valor total anual	Variação
2003	29379	
2004	30984	1605
2005	32215	1231
2006	33021	806
2007	34410	1389
2008	33470	-940
2009	34387	917
2010	34055	-332
2011	37125	3070
2012	37925	800
2013	38340	415
2014	39119	779
2015	39919	800
2016	41784	1865
2017	43569	1785

Tabela I.6: Diferenças das receitas em milhares de euros da Ponte 25 de Abril de 2003 a 2017

I.3 Aplicação dos Modelos da Teoria dos Valores Extremos

I.3.1 Modelo GEV Multivariado - Gráficos em detalhe

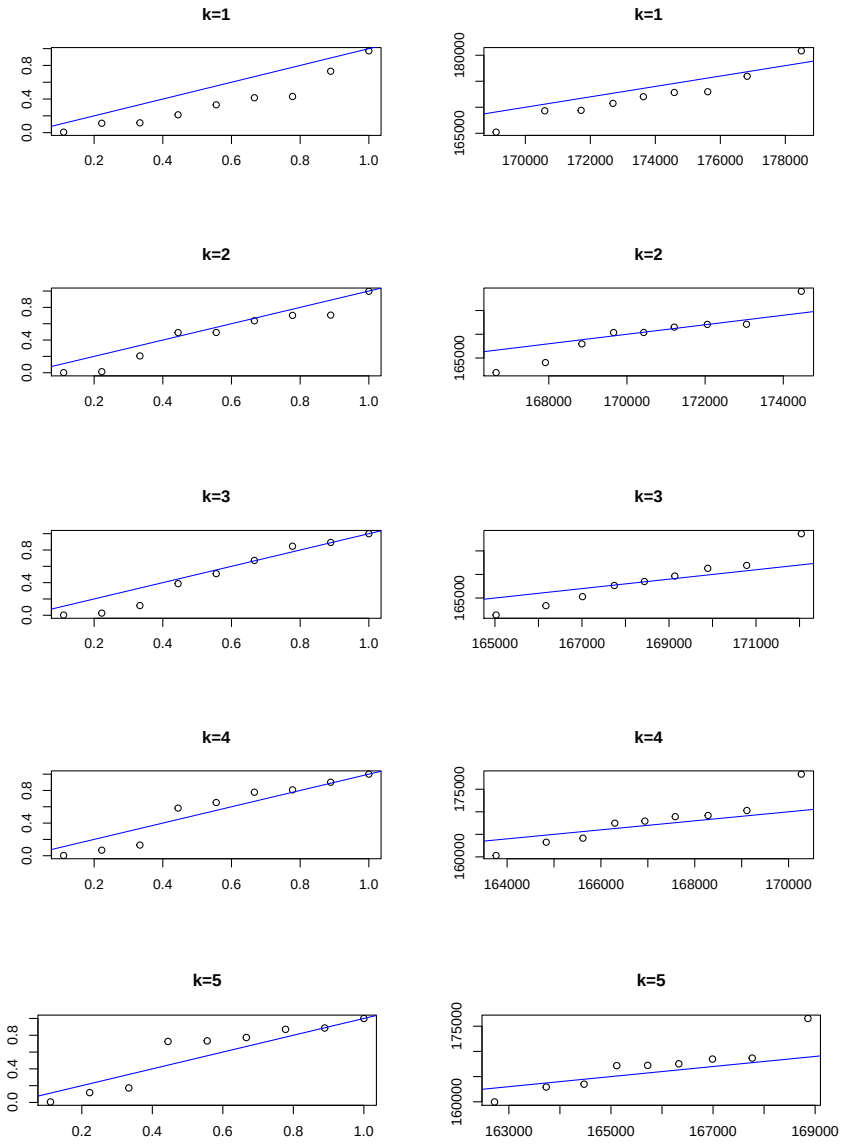


Figura I.3: Diagnóstico do modelo para os dados do tráfego da Ponte 25 de Abril com base no modelo ajustado da estatística das r maiores observações com $r = 5$. Gráficos de probabilidade (do lado esquerdo) e de quantis (do lado direito) para as estatísticas de k maiores observações, $k = 1, ..., 5$